Blind Dereverberation of Speech From Moving and Stationary Speakers Using Sequential Monte Carlo Methods

Christine Evers

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Edinburgh

September 2010

Blind Dereverberation of Speech From Moving and Stationary Speakers Using Sequential Monte Carlo Methods

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Edinburgh

Christine Evers
Room 2.01
Alexander Graham Bell Building
The King's Buildings
Mayfield Road
Edinburgh
EH9 3JL
Scotland, UK
c.evers@ed.ac.uk
Telephone: +44 (0)131 650 5655
Fax: +44 (0)131 650 6554

Last revision: September 2010

School of Engineering and Electronics
College of Science and Engineering
UNIVERSITY *of* EDINBURGH

Room 2.01
Alexander Graham Bell Building
The King's Buildings
Mayfield Road
Edinburgh
EH9 3JL
Scotland, UK
c.evers@ed.ac.uk
Telephone: +44 (0)131 650 5655
Fax: +44 (0)131 650 6554.

Major revision, September 2010.
Last printed revision with minor corrections, 22 October, 2010.

Typeset by the author with the LaTeX 2ε Documentation System, with AMS-LaTeX Extensions,
in 12/18 pt Times and Euler fonts.

INSTITUTE FOR DIGITAL COMMUNICATIONS,
School of Engineering and Electronics,
College of Science and Engineering,
Kings's Buildings,
Edinburgh, EH9 3JL. U.K.

Für Mama und Papa

"Out of clutter, find simplicity.
From discord, find harmony.
In the middle of difficulty lies opportunity."

ALBERT EINSTEIN, 1879-1955

Abstract

Speech signals radiated in confined spaces are subject to reverberation due to reflections of surrounding walls and obstacles. Reverberation leads to severe degradation of speech intelligibility and can be prohibitive for applications where speech is digitally recorded, such as audio conferencing or hearing aids. Dereverberation of speech is therefore an important field in speech enhancement.

Driven by consumer demand, blind speech dereverberation has become a popular field in the research community and has led to many interesting approaches in the literature. However, most existing methods are dictated by their underlying models and hence suffer from assumptions that constrain the approaches to specific subproblems of blind speech dereverberation. For example, many approaches limit the dereverberation to voiced speech sounds, leading to poor results for unvoiced speech. Few approaches tackle single-sensor blind speech dereverberation, and only a very limited subset allows for dereverberation of speech from moving speakers.

Therefore, the aim of this dissertation is the development of a flexible and extendible framework for blind speech dereverberation accommodating different speech sound types, single- or multiple sensor as well as stationary and moving speakers.

Bayesian methods benefit from – rather than being dictated by – appropriate model choices. Therefore, the problem of blind speech dereverberation is considered from a Bayesian perspective in this thesis. A generic sequential Monte Carlo approach accommodating a multitude of models for the speech production mechanism and room transfer function is consequently derived. In this approach both the anechoic source signal and reverberant channel are estimated using their optimal estimators by means of Rao-Blackwellisation of the state-space of unknown variables. The remaining model parameters are estimated using sequential importance resampling.

The proposed approach is implemented for two different speech production models for stationary speakers, demonstrating substantial reduction in reverberation for both unvoiced and voiced speech sounds. Furthermore, the channel model is extended to facilitate blind dereverberation of speech from moving speakers. Due to the structure of measurement model, single- as well as multi-microphone processing is facilitated, accommodating physically constrained scenarios where only a single sensor can be used as well as allowing for the exploitation of spatial diversity in scenarios where the physical size of microphone arrays is of no concern.

This dissertation is concluded with a survey of possible directions for future research, including the use of switching Markov source models, joint target tracking and enhancement, as well as an extension to subband processing for improved computational efficiency.

Declaration of Originality

I hereby declare that, except where otherwise stated in the text, the work described in this dissertation is composed and originated entirely by myself in the School of Engineering, Institute for Digital Communications, at the University of Edinburgh. It is not the result of work done in collaboration, and has not been submitted to any other University for any other degree. The length of this dissertation, including figures, tables, footnotes, appendices, and bibliography, is 269 pages and does not exceed $70,000$ words.

CHRISTINE EVERS
October 22, 2010

Acknowledgements

I would like to thank my supervisor, Dr James R. Hopgood for many technical discussions as well as the template file for the title page of this dissertation. I would also like to thank Dr Judith Bell, who was jointly supervising me during my first two years, for the knowledge transfer and technical advice. Thanks are also due to Dr Harald Haas for his guidance and continued interest in my academic progress, as well as to Nicola Ferguson for all her help and support in the past years.

A very special thanks goes to Dr Patrick A. Naylor and Dr Emanuel A. P. Habets for the amazing opportunity of presenting the tutorial on Speech Dereverberation with them at EUSIPCO 2010. I would also like to thank Dr Nikolay D. Gaubitch for providing me with the code for the Yegnanarayana implementation as well as the generalised discrete Fourier transform (GDFT) implementation to tinker around with.

Many thanks also to the Scottish Funding council for my position and the financial support in the Joint Research Institute in Signal & Image Processing (JRI-SIP) at the University of Edinburgh as part of the Edinburgh Research Partnership in Engineering and Mathematics (ERPem).

I would like to thank my friends for all the hours of ~~procrastination~~ fruitful discussions throughout the past three years. I would especially like to thank Richard for his patience, encouragement and support, especially over the last painstaking weeks of writing this dissertation.

Above all, I thank my mother, without whom this thesis would not have been possible or ever even been started. Thank you for always believing in me, for always supporting me, for always encouraging me, and for never losing your hope.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

AIC	Akaike's information criterion
AIR	acoustic impulse response
AR	autoregressive
ASR	automatic speech recognition
ARMA	autoregressive moving average
BIBO	bounded-input bounded-output
BS	block stationary
BSAR	block stationary AR
BSD	Bark spectral distance
CD	compact disc
cdf	cummulative distribution function
CGSS	conditionally Gaussian state-space
CONDENSATION	conditional density propagation
CRLB	Cramér-Rao lower-bound
DC	direct current
DFT	discrete Fourier transform
DSB	delay-and-sum beamformer
DYPSA	dynamic programming projected phase-slope algorithm
EM	expectation-maximization
FFT	Fast Fourier transform
FIR	finite impulse response
FT	Fourier transform
GCI	glottal closure instant

HERB harmonicity based dereverberation

HMM hidden Markov model

IFT inverse Fourier transform

i. i. d. independent and identically distributed

IIR infinite impulse response

ISM image-source method

JMS jump Markov system

LHS left hand side

LP linear prediction

LPC linear predictive coding

LSD log-spectral distortion

LTI linear time-invariant

LTV linear time-varying

MA moving average

MAP maximum a posteriori

MC Monte Carlo

MCMC Markov chain Monte Carlo

MDL minimum description length

MIMO multi-input, multi-output

ML maximum-likelihood

MM multiple model

MMSE minimum mean-square error

MS mean square

MSE mean squared error

NASA National Aeronautics and Space Administration

OM-LSA optimally-modified log spectral amplitude

PARCOR partial correlation

PCA principal component analysis

PDA personal digital assistant

PDE partial differential equation

pdf probability density function

PFS parallel formant synthesizer

PSD power spectral density

PR perfect reconstruction

RADAR radio detection and ranging

RBPF Rao-Blackwellized particle filter

RHS right hand side

RIR room impulse response

RMSE root mean squared error

RLS recursive least squares

RTF room transfer function

SIS sequential importance sampling

SIR sequential importance resampling

SLLN strong law of large numbers

SMC sequential Monte Carlo

SNR signal-to-noise ratio

SONAR sound navigation and ranging

SRR signal-to-reverberant component ratio

STFT short-time Fourier transform

SVD singular value decomposition

TIMIT Texas Instruments, Inc., and Massachusetts Institute of Technology

TVAP time-varying all-pole

TVAR time-varying AR

VAD voice activity detector

VB variational Bayes

WGN white Gaussian noise

Nomenclature

Operators

$*$	Convolution
$\mathbf{x}^{H}$, $\mathbf{X}^{H}$	Vector and matrix Hermitian transpose
$\mathbf{x}^{T}$, $\mathbf{X}^{T}$	Vector and matrix transpose
$\mathbf{x}^{\star}$	Complex conjugate
$\delta(x)$	Delta Dirac function
$\operatorname{diag}[\mathbf{x}]$	Diagonal matrix with elements of vector $\mathbf{x}$ on the diagonal
$\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})]$	Expected value
$\int_{\mathcal{X}} f(x)dx$	Integral of $f(x)$ over x within the region $\mathcal{X}$
$\mathbb{I}_{\mathcal{X}}(\mathbf{x})$	Indicator function over the set $\mathcal{X}$
$\partial/\partial T$	Partial derivative
$\mathcal{N}(\mathbf{x} \mid \mu, \Sigma)$	Multivariate Gaussian distribution
$\Pr(x)$	Probability
$\mathcal{U}[a, b]$	Uniform distribution between a and b
$\operatorname{var}_{p(\mathbf{x})}[\mathbf{x}]$	Variance
$\lvert x \rvert$	Absolute value
$x \sim y$	Sample x from y

Constants

K	Number of resonators
M	Number of subbands
N	Number of particles
N_{eff}	Effective sample size, providing measure of degeneracy based on importance weights
P	Model order of channel

Q	Model order of source
ω_s	Radial sampling frequency
f_s	Sampling frequency in cycles per sample

Matrices

$\boldsymbol{\varphi}_{0:t}$	Collection of unknown variables, $2Q + M(P+1) + 1 \times t$
$\boldsymbol{\theta}_{0:t}$	Collection of source model parameters, process and measurement noise log-variance; $Q + 1 + M \times t$
$\mathbf{z}_{0:t}$	Collection of source signal and channel parameters; $Q + MP \times t$
$\mathbf{a}_t$	Source parameters; $Q \times 1$
$\mathbf{A}_t$	Source transition matrix; $Q \times Q$
$\mathbf{b}$	Channel parameters; $MP \times 1$
$\mathbf{x}_t$	Current and past Q anechoic source signal samples; $Q \times 1$
$\mathbf{y}_t$	Echoic bservations at all sensors; $M \times 1$
$\boldsymbol{\Sigma}_{t\|t}$	Covariance of the corrected posterior probability density function (pdf) of $\mathbf{z}_t$; $MPQ \times MPQ$ matrix
$\boldsymbol{\Sigma}_{t\|t-1}$	Covariance of the predicted posterior pdf of $\mathbf{z}_t$; $MPQ \times MPQ$ matrix
$\boldsymbol{\Sigma}_{\mathbf{b},t}$	Covariance of channel posterior pdf; $MP \times MP$
$\boldsymbol{\Sigma}_{\mathbf{x}_{t\|t}}$	Covariance of the corrected source signal posterior pdf; $Q \times Q$ matrix
$\boldsymbol{\Sigma}_{\mathbf{x}_{t\|t-1}}$	Covariance of the predicted source signal posterior pdf; $Q \times Q$ matrix
$\boldsymbol{\mu}_{t\|t}$	Updated mean of the corrected posterior pdf of $\mathbf{z}_t$; $MPQ \times 1$
$\boldsymbol{\mu}_{t\|t-1}$	Predicted mean of the predicted posterior pdf of $\mathbf{z}_t$; $MPQ \times 1$
$\boldsymbol{\mu}_{\mathbf{x}_{t\|t}}$	Corrected mean of the source signal posterior pdf; $Q \times 1$
$\boldsymbol{\mu}_{\mathbf{x}_{t\|t-1}}$	Predicted mean of the source signal posterior pdf; $Q \times 1$
$\boldsymbol{\mu}_{\mathbf{b},t}$	Mean of channel posterior pdf; $MP \times MP$
$\boldsymbol{\Sigma}_{\mathbf{z}_{t,b}}$	Residual covariance of the channel
$\mathbf{K}_{\mathbf{b}_t}$	Optimal Kalman gain of the channel
$\mathbf{K}_t$	Optimal Kalman gain of $\mathbf{z}_t$
$\mathbf{Y}_{t-1}$	Matrix of past observations at all microphones; $M \times PM$ matrix
$\boldsymbol{\Sigma}_{z_t}$	Measurement residual covariance
$\mathbf{K}_{\mathbf{x}_t}$	Optimal Kalman gain of the source signal
$\boldsymbol{\Sigma}_{\mathbf{v}_t}$	Process noise covariance matrix; $Q \times Q$

$\mathbf{\Sigma}_{\mathbf{w}_t}$	Measurement noise covariance matrix over all microphones; $M \times M$
$\mathbf{\Sigma}_{(\mathbf{b}\|\mathbf{x})_t}$	Cross-correlation between the channel and source
$\mathbf{\Sigma}_{(\mathbf{x}\|\mathbf{b})_t}$	Cross-correlation between the source and channel

Probabilities

$p(\mathbf{y}_{1:t} \mid \mathbf{x}_{0:t})$	Likelihood
$p(\mathbf{x}_{0:t}, \boldsymbol{\theta}_{0:t} \mid \mathbf{y}_{1:t})$	Joint posterior pdf of $\mathbf{x}_{0:t}$ and $\boldsymbol{\theta}_{0:t}$
$p(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t})$	Posterior pdf
$p(\mathbf{y}_{1:t})$	Evidence
$p(\mathbf{x}_{0:t})$	Prior pdf
$\pi(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t})$	Proposal distribution; also known as importance function, or hypothesis distribution

Symbols

$\mathbf{x} = \{\mathbf{x}_\ell\}_{\ell \in \mathcal{L}}$	Set of elements of $\mathbf{x}$ in $\ell \in \mathcal{L}$
$\ell \in \mathcal{L}$	Set $\ell = 1, \ldots, L$
$\mathbf{I}_L$	Identity matrix of size X
$\mathbb{N}$	Set of integers
$\mathbb{R}^\ell$	Set of real numbers of dimension ℓ

Units

dB	Decibel
Hz	Hertz
m	Meter
s	Second

Variables

$G(z)$	Glottal pulse spectrum
$H(z)$	room transfer function (RTF)
$N(z)$	Noise spectrum
$R(z)$	Radiation model
$V(z)$	Vocal tract transfer function
v_t	Excitation sequence of source signal
w_t	Measurement / observation noise sequence
$X(z)$	Speech spectrum
x_t	Unobserved anechoic source signal
y_t	Observed echoic signal

$\hat{x}_t$	Source signal estimate
$\{a_{q,t}\}_{q\in\mathcal{Q}}$	Source parameters
$\{b_{p,t}\}_{p\in\mathcal{P}}$	Channel parameters
ϕ_{v_t}	Log-variance of the process noise
ϕ_{w_t}	Log-variance of the measurement noise
$\sigma^2_{v_t}$	Process noise / excitation variance
$\sigma^2_{w_t}$	Measurement noise variance
$\{\lambda_{q,t}\}_{q\in\mathcal{Q}}$	Forward lattice stage output
$\{\phi_{q,t}\}_{q\in\mathcal{Q}}$	Backward lattice stage output
$\{\psi_{q,t}\}_{q\in\mathcal{Q}}$	Reflection / PARCOR coefficients
$\{A_k\}_{k\in\mathcal{K}}$	Cross-sectional area of tube $k \in \mathcal{K}$ in the acoustic tube model of the vocal tract
$\{B_{k,t}\}_{k\in\mathcal{K}}$	Resonator 3dB bandwidth
$\{f_{k,t}\}_{k\in\mathcal{K}}$	Resonator frequency
$\{g_{k,t}\}_{k\in\mathcal{K}}$	Resonator gain
$r_{k,t}$	Reflection coefficients of tube $k \in \mathcal{K}$ in the acoustic tube model of the vocal tract
$\tilde{w}_t$	Re-normalised importance weights for particle i
w_t	Un-normalised importance weights; do not include evidence term in ratio between posterior and importance density
$w^\star_t$	Normalised importance weights; ratio / discrepancy between posterior and importance density

Part I

Thesis, introduction, and motivation

Chapter 1

Introduction

1.1 Motivation

Speech is the most profound, natural, and efficient form of human-to-human communication. A large proportion of human-to-human communication takes place in offices, living rooms, theatres, cathedrals, or other confined communal environments. Thus, speech is subjected to reflections of the sound wave off surrounding walls and obstacles, such as the floor, ceiling, walls and neighbouring objects. If the reflections arrive within short time intervals of the order milliseconds after the direct-path signal, the delayed reflections are referred to as reverberation. In contrast, echoes are sound reflections that are noticeable as a spatially or temporally separated repetition of the direct path signal. Reverberation was shown to be equally disturbing as an echo [4]. Morevover, as reflections are not temporarily distinct, dereverberation, i.e., the removal of reverberant effects, is an inherently difficult problem.

1.2 The distorting effects of reverberation

The reduced intelligibility of speech in reverberant environments can be attributed to several factors. As proposed by Nabalek *et al.* [5], self-masking and overlap-masking are two major contributors to impaired intelligibility under reverberation. Self-masking is the alteration of phonemes in time and frequency caused by the distortion of sound onsets and decays of transient sounds [6]. Transient sound bursts become less abrupt [7] and formant transitions between vowels are disrupted due to temporal smearing. Overlap-masking is referred to the masking of phonemes by preceding phonemes and their reflections [6], leading to impoverishment of phonemes. Furthermore, reverberation reduces the perception of consonants, an effect referred to as articulation loss of consonants by Peutz [8]. According to Peutz, the articulation loss of phonemes is proportional to the source-sensor distance and reverberation time. The distortion of

consonants is more detrimental than impaired vowels as the intelligibility of speech is highly dependent on the clear perception of consonants.

Although the human sound perception mechanism often copes with reverberation due to well-trained psychoacoustic processing, e.g., by exploitation of spatial perception or partial lip reading, digitally recorded reverberant speech exhibits significant distortions. Applications affected by reverberation include the following examples:

- Audio conferencing applications facilitate live interaction between groups of people remote from each other by means of one or several locally central microphones in a room, recording and transmitting the mass articulation to the remote end.
- Hearing aids use microphone arrays improve hearing capability by filtering and amplifying sound signals of interest to the listener.
- Handsfree devices, such as mobile phones on speaker mode, personal digital assistants (PDAs), or Skype phones record speech signals at a distance, e.g., mounted on the dashboard of a car or lying on the desk of a workstation.
- Naval applications using sound navigation and ranging (SONAR) [9], or airborne applications using radio detection and ranging (RADAR), where a propagated signal is scattered through obstructions and undesired objects.

Dereverberation, i.e., the enhancement of the reverberant signals to obtain an estimate of the anechoic speech signal, is thus crucial for the quality and intelligibility of speech for a vast number of signal processing and scene analysis applications, including teleconferencing, automatic speech recognition, hearing aids, or source localisation, tracking and identification.

1.3 Removing the effects of reverberation

Driven by consumer demand, speech dereverberation has become a prominent topic in the research community over the last few years. As exact prior knowledge of the anechoic speech signal, the reverberant room response, or noisy events such as opening and closing of doors, is generally unavailable in practice, speech dereverberation needs to be performed blindly. Blind speech dereverberation techniques in the literature are roughly divided into spatial filtering techniques, homomorphic transformations, spectral enhancement, approaches exploiting explicit speech modelling and linear prediction (LP) residual enhancement.

Valuable insight can be gained from the concepts behind the algorithms in the literature. For example, harmonicity based dereverberation (HERB) exploits the assumption that voiced speech is harmonic and speech can thus be dereverberated by

the extracting harmonic components and constructing a filter that suppresses reverberation by comparing the discrepancy between the extracted harmonic spectrum and the reverberant spectrum. Explicitly embracing the properties of voiced and unvoiced speech as prior information in dereverberation frameworks can significantly improve estimation of the underlying anechoic speech signal and is therefore highly appealing. LP residual enhancement techniques assume that the RIR leaves the source production model unchanged. Dereverberation is attempted by iteratively estimating the speech source assuming the reverberant channel is constant and subsequently estimating the channel assuming the source is constant. Rather than attempting to simultaneously estimate (and possibly optimise) the source and channel models, successive updates of two separate estimators with their respective outputs is a compelling notion.

Almost all blind dereverberation approaches in the literature are founded on rigid assumptions about either the dynamics of the source signal, or properties of the speech production and reverberant channel. The quality of reverberant speech can be improved in scenarios where the conditions of the respective approaches are satisfied. However, as entire dereverberation algorithms are in fact derived from the model assumptions, extensions to different scenarios either lead to inferior dereverberation performance or are prohibited altogether. In particular, apart from the odd exception, dereverberation algorithms rarely attempt to account for moving speakers, where the source-sensor position and hence the RTF changes with time.

1.4 Aim of this dissertation

This dissertation considers the rigorous constraints to very specific aspects of the overall problem of speech dereverberation as the actual deficiency of the majority of approaches in the literature. Therefore, the aim of this dissertation is to develop a speech dereverberation framework that relies on very generic assumptions about the speech production mechanism and reverberant environment. The developed framework therefore facilitates extensions to different speech sounds, such as voiced, unvoiced, transient, or fricative sounds, stationary and moving speakers.

Therefore, rather than attempting to outperform approaches in the literature in terms of the reconstruction of the anechoic source signal, the aim of this dissertation is to develop a flexible algorithm that avoids the restriction of blind speech dereverberation to specific sub-problems.

The framework can be targeted at specific problems by choosing appropriate models. For instance, the dereverberation of voiced and unvoiced speech can be tackled

by modelling the source production mechanism by source models accounting for the harmonicity of voiced speech and the turbulent noise resemblance of unvoiced speech. Furthermore, dereverberation can be extended to moving speakers by choosing time-varaing channel models in order to capture the change of the RIR with varying source-sensor positions.

Ultimately, the incorporation of models specifying underlying signal and model structures is equivalent to the inclusion of prior knowledge about the speech production and reverberation system. Rather than exclusively relying on the reverberant observations, *belief* about the nature of speech and reverberation is embraced as well. The inclusion of prior knowledge in addition to information inferred from the observations is at the very core of Bayesian estimation techniques: the prior pdf of any unknown and desired variables and the likelihood function of the reverberant data are combined to form the posterior pdf using Bayes's theorem. Bayesian estimators maximise the posterior pdf or minimise the posterior expected value of a loss function to obtain estimates of the unknown parameters.

Therefore, this dissertation investigates whether a stochastic Bayesian perspective on blind speech dereverberation can provide added benefit of generality and flexibility to the field.

1.5 Bayesian estimation

In 1961, the National Aeronautics and Space Administration (NASA) launched its Apollo program, culminating eight years later in the first lunar landing in human history. At the very core of the navigation computer of Apollo 11 resided an analytical recursive estimator known as the Kalman filter that NASA developed in collaboration with Rudolf Kalman. The Kalman filter obtains an estimate of source signals buried in noise by sequentially predicting and updating the source signal posterior pdf (see, e.g., [10–15] for a general overview and [16–18] from a Bayesian perspective). Being the optimal estimator for linear systems and easily extendible to a variety of problems, the Kalman filter is a highly popular estimator to this date, almost half a century after the lunar landing. Many estimators therefore strive to utilise the Kalman filters for the estimation of source signals for its optimal properties and ease of implementation.

In order to obtain an estimate of the source signal, Kalman filters require knowledge of any underlying model parameters. However, due to the blind nature of dereverberation problems, prior knowledge of the speech production and channel model parameters is not available in practice. Furthermore, as realistic data is highly com-

plex, of high dimensionality, and possibly subject to non-linearities, direct analytic estimation using the Kalman filter is not possible as the posterior pdf is analytically intractable.

Many approximate solutions for non-linear Kalman filtering have been proposed over the years, e.g., the extended Kalman filter, Gaussian sum approximations, or grid-based filters have been proposed as an alternative of the Kalman filter for non-linear, non-Gaussian problems [19–21]. However, these approaches often ignore underlying statistical features of the signals, or are too computationally expensive for practical application in high-dimensional problems.

In the 1950s, the basic idea of sequential Monte Carlo (SMC) in the form of sequential importance sampling (SIS) emerged [22]. In order to approximate the intractable posterior pdf, a large number of random variates is drawn sequentially from a hypothesis distribution that approximates and has the same region support as the posterior pdf. However, due to the availability of comparatively modest computational power, SIS was largely disregarded for decades until its rediscovery in the early 1990s with the emergence of the resampling step [23] and the associated sequential importance resampling (SIR) approach. Similar to the Kalman filter, estimates are propagated recursively in time by construction of a prediction based on the past estimate trajectory (referred to as importance sampling), and correction the estimates using knowledge inferred from the measured data (known as the resampling step). SIR varies from the analytic Kalman filters in that they are numerical rather than analytic solutions. For instance, particle filters perform SMC estimation based on point-mass representations of pdfs: a large number of samples is drawn from a hypothesis distribution in order to approximate analytically intractable posterior pdfs.

Since the evolution of SIR, SMC swiftly became one of the most popular and active research areas of Bayesian estimation. Publications such as [2, 16, 17, 24, 25] stirred the interest in SMC approaches in the research community.

A particularly interesting development proposed in the mid-1990s is the concept of Rao-Blackwellisation of the sampling scheme for variance reduction in the estimates [26]. The state-space is partitioned into analytically tractable and intractable variables. Whilst the intractable parameters are estimated using SIR or variants thereof, analytically tractable substructures are exploited to estimate the corresponding parameters using their optimal estimators. Estimation of the analytically tractable substructures can often be achieved using the Kalman filter.

The optimal Kalman filter can thus be incorporated for state estimation in non-linear, non-Gaussian systems, where knowledge of the model parameters is not available *a priori*. In other words, optimal estimators can be utilised for blind non-linear estimation problems despite their restriction to non-blind and analytically tractable systems. Due to its Bayesian foundation, Rao-Blackwellised SMC thus allows for the combination of prior information inferred from models with the knowledge inferred from the measured data. Last but not least, Rao-Blackwellised SMC can be specified as generic algorithms valid for any choice of models facilitating sequential processing. Therefore, Rao-Blackwellised SMC algorithms are not *dictated* by models, but rather *benefit* from information inferred from carefully chosen models.

Rao-Blackwellised SMC therefore is a strikingly attractive option for the development of a flexible and extendible framework for blind speech dereverberation.

1.6 Thesis, contributions, and overview

The essential thesis of this dissertation is therefore that fundamental contributions towards the solution of the overall blind speech dereverberation problem can be made by the development of a flexible and extendible blind dereverberation framework that can be established from a Bayesian perspective using Rao-Blackwellised SMC. To substantiate this thesis:

Chap. 2 presents a critical survey of blind speech dereverberation approaches in the literature. Valuable insights, shortcomings, and limitations are assessed whilst emphasising how the insight can be utilised in the context of this dissertation. A related publication is [1] as below.

Contributed publications on the speech dereverberation literature

[1] P. Naylor, C. Evers, and E. A. P. Habets. Speech dereverberation. In *Proc. EUSIPCO*, Aalborg, Denmark, August 2010. Tutorial, proposal submitted for review.

Although Bayesian approaches benefit from appropriately chosen models, unsuitable models can be detrimental to even sophisticated algorithms.

Chap. 3 therefore establishes the necessary background knowledge of speech models required for the developments in Chaps. 7 and 8. Aptness at modelling different speech sounds and limitations are discussed in order to ensure appropriate implementation.

Chap. 4 studies the benefits and shortcomings for the models of the RTF for dereverberation of speech from stationary and moving speakers required in Chaps. 7 and 9.

Chap. 5 analyses the fundamental concepts for Rao-Blackwellised SMC approaches using Kalman filters for estimation of the analytically tractable substructures. The information presented in this chapter is heavily used in Chap. 6, where:

Chap. 6 proposes the novel blind speech dereverberation algorithm. Based on a general source and channel model, the minimum mean-square error (MMSE) estimator of the anechoic source signal, reverberant channel parameters, and speech production model is derived. The results reveal that the joint MMSE estimator *marginalises* both the anechoic source signal and channel parameters from the speech production model. In other words, the estimator is Rao-Blackwellised by isolating the anechoic speech signal and channel parameters from the parameter space of the speech production model. As the speech production model parameters are generally highly non-linear in the observations, analytical evaluation is not possible. Therefore, the source model parameters are estimated using SIR instead. Moreover, it is shown that the channel parameters are also marginalised from the speech signal. Uncertainty introduced by channel estimation is therefore incorporated in the speech signal estimate. Subsequently, both the source signal and channel parameters can be estimated using their optimal estimator which is of the form of a Kalman filter.

Channel estimation within SIR frameworks can be problematic if estimates are obtained via importance sampling. If the source-sensor positions are invariable, the acoustic channel does not vary with their position, such that the channel parameters are static and time-invariant. However, SIR implicitly assumes a dynamic on the underlying system parameters in order to recursively sample random variates. Importance sampling of static parameters therefore leads to poor channel estimates [27]. As the channel parameters are estimated using their optimal Kalman estimator in this framework, issues of static importance sampling are avoided.

Therefore, blind speech dereverberation is proposed by sequentially 1. importance sampling a large number of realisations (or particles) of the speech production model parameters, 2. for each resulting particle estimating the channel parameters, and 3. estimating the anechoic speech signal for each realisation of the speech production and channel parameters . Related publications are [2–5] as below.

Contributed publications on the proposed algorithm

[2] C. Evers and J. R. Hopgood, "Multichannel online blind speech dereverberation with marginalization of static observation parameters in a Rao-Blackwellized particle filter," *Springer J. Signal Process. Systems*, 2009, in print.

[3] C. Evers, J. R. Hopgood, and J. Bell, "Acoustic models for online blind source dereverberation using sequential monte carlo methods," in *Proc. IEEE Conf. ICASSP*, Las Vegas, NV, 24 Mar. - 4 Apr. 2008.

[4] ——, "Blind speech dereverberation using batch and sequential monte carlo methods," in *Proc. IEEE Conf. ISCAS*, Seattle, WA, 18-21 May 2008, invited paper.

[5] C. Evers and J. R. Hopgood, "Marginalization of static observation parameters in a Rao-Blackwellized particle filter with application to sequential blind speech dereverberation," in *Proc. EUSIPCO*, Glasgow, UK, Aug. 2009.

Chap. 7 applies a dynamic time-varying AR (TVAR) source model to the Rao-Blackwellized particle filter (RBPF). The source parameters vary according to a random walk constrained to poles within the unit circle to enforce stability of the process. Experiments using synthetic and speech data indicate significant improvement in speech quality of the estimated signal over the reverberant observations for stop consonants and fricative sounds. Related publications are [2–5] as above.

Chap. 8 extends the dynamic TVAR parameter model in Chap. 7 to a novel speech model based on parallel formant synthesizers (PFSs) reparameterised in terms of PARCOR coefficients in order to improve upon dereverbation of vowels. Results show overall improvement over the TVAR parameter model, and significant improvements for vowels. A related publications is [6] as below.

Contributed publications on the PFS model

[6] C. Evers and J. R. Hopgood, "Articulatory based speech models for blind speech dereverberation using sequential monte carlo methods," in *Proc. EUSIPCO*, Aalborg, Denmark, Aug. 2010, invited paper, submitted for review.

Chap. 9 extends the channel model from a stationary to moving speakers. An investigation of the RIR for changing source-sensor positions indicates that the channel parameters characterising the RIR vary with time. Furthermore, it is shown that the parameters can be approximated using Fourier polynomials. The channel is therefore modelled as a linear combination of channel parameters with Fourier basis functions. Results show that signal improvement can be achieved for sufficiently high source signal power. Related publications are given in [7–10].

Contributed publications on moving speakers

[7] C. Evers and J. R. Hopgood, "Parametric models for single-channel blind dereverberation of speech from a moving speaker," *IET J. Signal Process.*, vol. 2, no. 2, pp. 59–74, Jun. 2008.

[8] J. R. Hopgood, C. Evers, and S. Fortune, "Bayesian single channel blind dereverberation of speech from a moving speaker," in *Speech dereverberation*, P. A. Naylor and N. Gaubitch, Eds. Springer, 2010.

[9] J. R. Hopgood and C. Evers, "Towards single-channel blind dereverberation of speech from a moving speaker," in *IMA Intl. Conf. Math. Sig. Proc.*, Dec. 2006.

[10] ——, "Block-based TVAR models for single-channel blind dereverberation of speech from a moving speaker," in *Proc. IEEE Conf. SSP*, Madison, WI, 2007, pp. 274–277.

Chap. 10 evaluates the computational complexity of the RBPF. Results indicate that the rate of growth increases quadratically in the channel order and at least quadratically in the number of sensors. It is proposed that subband filtering approaches can be used to alleviate the computational burden.

Chap. 11 concludes the dissertation by summarising the results and issues discussed and encountered in these investigations and outlines directions future research could take. Specific details are given for a hybrid switching model combining the speech models in Chaps. 7 and 8, and for the inclusion of individual gain factors for multi-sensor processing.

Chapter 2

Critical overview of blind dereverberation approaches in the literature

Dereverberation techniques can be broadly divided into two categories: 1. reverberation suppression and 2. reverberation cancellation techniques.Reverberation suppression amplifies characteristics properties of the source signal whilst attenuating effects due to reverberation. Reverberation cancellation estimates the room acoustic transfer function used for equalisation of the distorting channel from the observed signal.

The aim of this chapter is to identify the advantages and drawbacks of blind dereverberation approaches in the literature in order to identify how the dereverberation algorithm in this dissertation sets itself apart, and extract beneficial features that can be incorporated in the proposed approach in this dissertation.

As this chapter is primarily concerned with the differences in *methodology*, the approaches presented in this discussion are grouped into spatial filtering (sect. §2.1), homomorphic transformations (sect. §2.2), spectral enhancement (sect. §2.3), explicit source modelling (sect. §2.4), and linear prediction (LP) residual enhancement (sect. §2.5). Conclusions are presented in sect. §2.6.

2.1 Blind dereverberation by spatial filtering

Spatial enhancement techniques are used for multi-microphone dereverberation. By attenuating signals from certain directions, the observed signal can be enhanced to spatially separate reverberant paths from the direct path signal. The simplest form of spatial enhancement is obtained by delay and sum beamforming: a microphone array increases the sensitivity of the sensors in the direction of the audio source, such that

sensitivity in the direction of interfering sources is decreased. Interference can thus be removed even on the same frequency band as the source signal. Incoming plane waves of spatially distinct sources arrive at the sensors of a microphone array with a slight time delay. Delay-and-sum beamformers introduce suitable delays by means of weights to each channel and linearly combine the set of the received signals at the sensor array. Physically speaking, a beam is formed in the direction of the source by constructively adding and thus amplifying the coherent direct path component across channels, whilst smearing with time and thus attenuating interference such as reverberation, exhibiting an incoherent structure after the filtering. The weights determine the filtering characteristics of the beamformer and can be chosen to provide a fixed response independent of the received data (data independent beamformers), or to optimise the beamformer response based on statistics of the received data (statistically optimum beamformers). Beamforming is therefore equivalent to temporal finite impulse response (FIR) filter and hence often referred to as spatial filtering [28, 29]. A broad overview of beamforming techniques is given by Van Veen and Buckley in [28].

An investigation of the direct to reverberant ratio of delay and sum beamformers applied for reverberation [30] shows that the relative improvement depends on the microphone spacing and the distance of the source from the array, but is independent of the reverberation time. Nonetheless, as the frequency increases, the beam shape narrows. Further, the frequency response depends on the angle of arrival, such that signals arriving off-axis will be subject to spectral colouration by the array.

Several extensions to beamforming have been investigated: in a sub-band extension of the delay-and-sum beamformer, Allen *et al.* [31] exploit the observation that the tail of the room impulse response (RIR) is mostly uncorrelated. Within each sub-band, the delay between coherent parts, corresponding to the source signal and early reflections, is removed by phase shifting the sub-band signals. In order to amplify the source signal and early reflections, the resulting sub-band signals are added and the sub-band gains are adjusted by the cross-correlation between the distorted observed signals. Frequency bands with low levels of coherence are thus attenuated whilst highly coherent bands containing the direct path component and early reflections are amplified.

Flanagan *et al.* [32] quantise the spatial range of the beamformer into overlapping regions which are scanned sequentially, creating a matched filter beamformer. By utilising speech characteristics, such as periodicity of speech at the vocal cord rate, as well as accumulation of acoustic energy primarily in voiced sounds, and the tendency to be bursty in time, the sound source can be distinguished from continuous background

noise by spectral analysis of the beamformer output.

In environments with continually changing characteristics, using a predetermined set of weights might not be suitable. Adaptive beamformers automatically adjust the weights [28, 29]. In cases of lack of knowledge about the desired signal, linear constraints can be introduced in the adaption rule in order to allow for control over the adapted response of the beamformer [33]. However, adaptive beamformers are susceptible to changing RIRs, such that most adaptive beamforming techniques for the enhancement of speech become ineffective in reverberant environments.

[29, 31–34] thus attempt to remove interference by forming a beam in the direction of the source and thus removing interference with different directionality. However, reverberant reflections arrive from a multitude of different positions in the room and are thus highly likely to interfere with the beam path. Hence, the beamforming techniques described above only removes reverberation to some degree. To improve dereverberation using beamforming, Affes and Grenier [35] propose a matched filter beamforming approach that adaptively estimates the channel response and convolves the received signals with the inverse of the resulting RIR. Flanagan *et al.* extend their approach in [32] to a three-dimensional array [34], forming additional beams steered in the direction of strong initial reflections. Similarly to the image-source method (ISM) (see sect. §4.3) [31], the reflections are considered as mirror sources [36]. However, both [35] and [34] require at least partial knowledge of the RIR which is generally not available in practice.

2.1.1 Issues and insights of beamforming techniques

Spatial enhancement techniques generally depend on beamforming techniques, where microphone arrays are steered in the direction of the desired speech source. Spatial diversity can therefore be exploited to infer knowledge about reverberant environment and hence to recover the source signal. However, as reflections arrive from a multitude of directions and are likely to interfere with the beam path. Furthermore, adaptive beamformers are susceptible to changes in the RIR. Extensions to beamforming using matched filters, or by steering in the direction of strong initial reflections, require partial knowledge of the RIR, and are hence inappropriate for blind speech dereverberation.

2.2 Homomorphic transformation

Homomorphic transformations reflect non-additively combined signals to a domain where they can be considered as additively mixed. The domain of homomorphic

transformations is called the *cepstral* domain, an anagram derived from the word "spectrum" in order to distinguish from the frequency domain. Further terminological replacements that are utilised in the research community are "liftering" instead of "filtering" and "quefrency" instead of "frequency".

The cepstrum of a signal is given by

$$y_t^c = \text{IFFT}\left[\ln \text{FFT}\left[y_t\right]\right] \tag{2.1}$$

where FFT and IFFT denote the Fast Fourier transform (FFT) and its inverse respectively, y_t is the observed signal sample at time $t \geq 0$ and y_t^c is the signal cepstrum. Convolution in the time domain is equivalent to addition in the cepstrum domain. Thus, the source signal can be recovered from its observations using linear filtering techniques rather than by deconvolution of the source signal from the RIR.

In order to separate the speech cepstrum from that of the reverberations, it is observed that the speech cepstrum typically only has cepstral components concentrated around the cepstrum origin, whilst the RIR cepstrum is characterised by pulses with ripples located far away from the origin [37]. Thus, by suppressing any high cepstral components and maintaining low-time components only, the speech signal can be extracted from the distorted cepstrum. This technique is also known as *liftering* and is essentially a low-pass filter, i.e., a multiplication of the cepstral signal by a window in the cepstral domain [38, 39]. However, in order to perform liftering, a cutoff threshold is required, which is often acquired using empirical evidence. Heuristic measures of cutoff times lead to inconsistent dereverberation performance and introduce additional distortions due to framing effects of the threshold window.

Since the vocal tract model usually varies faster than the RIR, the channel response can be assumed to be approximately stationary over short time frames. Assuming that speech signals are zero-mean, whilst the mean of the room transfer function (RTF) is non-zero, the mean of the reverberant signal can be extracted by averaging over small time frames of the observed signal. A clean speech estimate can thus be obtained by cepstral subtraction of the mean of the observed signal over short time frames from the observed cepstrum. Effectively, the direct current (DC) component of the RIR is removed retaining only the time-varying components of speech [40]. However, cepstral mean subtraction implicitly assume time-invariant channels and require zero-mean speech signals.

Homomorphic transformations are mostly applied for dereverberation in speech

recognition systems, see, e.g., [38, 41–43], but have also been successfully applied for the restoration of archived records in order to remove surges in volume due as the pitch of the voice reaches the resonances of the recording equipment [44]. Restoration using the homomorphic approach is reported to remove megaphone effects whilst retaining the acoustic flavour of the recording [45].

2.2.1 Issues and insights of homomorphic transformations

Cepstral approaches have been reported to work effectively for speech recognition systems. Nonetheless, late reverberation can overlap in cepstrum domain with peaks of the speech signal, prohibiting disambiguation and identification of the clean source signal for speech estimation systems. Cepstral mean subtraction in these cases leads to artificially sounding signal estimates, rendering cepstral transforms for dereverberation of speech in highly reverberant rooms inappropriate for speech communications systems where the *human* auditory system is targeted.

Nonetheless, homomorphic transformations offer interesting insight into the separability of speech signals and reverberant distortions, laying the foundations for spectral enhancement techniques (see sect. §2.3). Notably, a very similar principle to homomorphic transformations can be applied to dereverberation of stationary speakers in the time domain using histograms: human speech is a highly time-varying function (see sect. §3.3), whereas the reverberant channel for a stationary speaker can be considered static if the room transfer function is not physically altered, i.e., doors / windows are not open / closed, furniture not moved, etc. Furthermore, both the RTF as well as the human speech production system can be modelled using all-pole models (see sects. §3.3 and §4.4). The stationarity of the channel vs. non-stationarity of speech can therefore be exploited by extracting the autoregressive (AR) parameters of reverberant speech signals over a sliding window and computing the histogram over the poles, i.e., the roots of the AR parameters. Whilst the poles characterising the time-varying speech spectrum smear over the floor of the pole histogram, the stationary channel poles appear as distinct peaks in the histogram. Using clustering methods, the reverberant channel can thus be identified from the pole histogram and represent an estimate of the reverberant channel modelled as an all-pole filter. Using inverse filtering (or channel equalisation the reverberant channel effects can thus be removed from the observations, yielding an estimate of the clean speech signal.

Whilst the histogram method offers viable identification of low-order channel models for stationary speakers, realistic channel models usually involve several hundred parameters to accurately reflect the reverberant properties (see sect. §4.4). For high model orders, clustering can prove problematic due to closely spaced pole positions

whose peaks potentially merge, hence prohibiting exact peak identification.

Based on the insight gained from the histogram method, a more advanced Bayesian approach in [1] as discussed in sect. §2.4 exploits the non-stationarity of speech signals in order to estimate speech signals generated using a linear time-varying (LTV) speech production model and distorted by linear time-invariant (LTI) channel models. Based on an iterative batch approach, the algorithm firstly estimates the source model assuming the channel as a nuisance parameter, and secondly estimating the channel assuming the source as a nuisance parameter. Using the source and channel model, the speech signal can be recovered from the reverberant observations by inverse filtering with the channel filter.

In a more insightful manner, the speech dereverberation algorithm proposed in Chap. 6 operates in a similar way by 1. sequentially estimating the source model parameters based on the observations only, 2. using the source parameter estimates and observations obtaining an estimate of the parameters of an LTI channel (extendible to LTV channel models as discussed in Chap. 9) using its optimal estimator, and 3. estimating the source signal using its optimal estimator based on the channel and source parameter estimate.

2.3 Spectral enhancement

Spectral enhancement techniques for speech dereverberation modify the short-time spectrum of the reverberant observed signal. The first spectral enhancement approach for speech dereverberation was developed as a side-product of an extension to spatial filtering, and is the sub-band beamformer proposed by Allen *et al.* [31] as discussed in sect. §2.1.

Spectral enhancement techniques for noisy speech have been a popular topic in the literature since Allen's pioneering paper in 1977. A statistical approach minimising the mean squared error (MSE) of a distortion measure between the clean and estimated signal is proposed, for example, in [46]. Juang and Rabiner [47] use hidden Markov processes, where the probability densities of the speech and noise processes are first estimated using long anechoic speech and noise training sequences. The estimates are used to derive an estimator of the source signal. Ephraim and van Trees [48] propose a subspace approach decomposing the vector space of the noisy signal into a signal-plus-noise and a noise subspace, removing the latter to estimate the source signal. An overview of spectral enhancement techniques can be found in, e.g., [49] and references

therein.

However, spectral enhancement for speech dereverberation has evolved as an active field of research with the beginning of the new millenium: Lebart *et al.* [50] propose an approach to single microphone dereverberation by spectral subtraction of the late reverberation energy from the spectrum of the observations. Late reverberant energy is estimated using a statistical model of the RIR, specified as Gaussian noise modulated by an exponentially decaying function whose decay is governed by the reverberation time. Based on an estimate of the late reverberant energy, the spectral attenuation factor in the spectrum of the observations due to late reverberations can be established [51]. An estimate of the speech signal and early reflections is thus obtained by filtering the spectrum of the distorted observations with the spectral attenuation factor. Spectral subtraction therefore attenuates late reverberations, whilst amplifying the spectrum of the speech signal and early reflections.

The approach was extended to multiple microphones by Habets in [52]. Wen *et al.* [53] compared the dereverberation performance of the approach in [52] with that of a delay-and-sum beamformer (DSB) in terms of subjective and objective measures, indicating an improved from approximately a Bark spectral distance (BSD) of 0.02 sones for the DSB to a BSD of 0.05 sones for multi-microphone spectral subtraction for a reverberation time of $T_{60} = 0.291$s.

Although spectral subtraction can be successfully applied to blind dereverberation problems, interference cause by noise sources in addition to reverberation is not considered in either [50] or [52]. The optimally-modified log spectral amplitude (OM-LSA) estimator was therefore proposed [54] as an extension of the spectral attenuation factor in [50] to recover speech signals buried in reverberation and noise, leading to improved BSD scores as compared to spectral subtraction. The OM-LSA is a spectral gain function that minimises the mean-square error of the log-spectral amplitude of speech and its estimate and is determined by statistical models of speech presence and absence.

Wu and Wang in [55] propose a combination of spectral subtraction and LP residual enhancement for speech dereverberation: In a first step, the LP residual enhancement technique in [56] discussed in sect. §2.5 is used to enhance the ratio of the direct signal to the reverberant signal (see sect. §2.5). Late reverberation is removed in a second step by application of spectral subtraction.

The sequential algorithm proposed by Yoshioka *et al.* [57] is one of the few blind

dereverberation approaches accounting for changes in the speaker location. Operating in the short-time Fourier transform (STFT) domain, the static channel parameters are recursively updated using a Bayesian recursive least squares (RLS) algorithm. The proposed RLS algorithm can be written in the form of the update equations of a Kalman filter as derived in Appendix A.1 (see sect. §5.3 for a detailed discussion of Kalman filters), although the authors have not explicitly commented on the fact. Given the channel estimates, the the STFT spectrum of the source signal is reconstructed by filtering the spectrum of the observed signal with the channel estimate. Results indicate that a distance measure of the estimated signal requires approximately 5s of data, i.e., $80,000$ samples at a sampling rate of 16kHz, for convergence towards a steady state for each change in the speaker position. It is argued that the convergence time can be reduced by the incorporation of prior knowledge about the RIR.

2.3.1 Issues and insights of spectral enhancement

Most spectral enhancement approaches for speech dereverberation are based on spectral *subtraction* techniques, removing the effects of late reverberations. Retaining early reflections in the estimated signal is desirable in scenarios where speech is digitally recorded and the enhanced estimate is played back in almost anechoic environments, e.g., via headphones or hearing aids. Reinforcement of the direct path signal by early reflections is thus facilitated that would otherwise not be available due to the lack of an echoic surrounding environment.

However, in many applications, speech is recorded in highly reverberant rooms and played back using loudspeakers. For instance, during conference calls it is desirable to dereverberate the signal from a remote party and play it back to the participants in the local office. As the remotely recorded speech signal is radiated locally in an echoic environment, early reflections are naturally introduced by means of the local RIR. It is therefore desirable to remove both early and late reflections and reconstruct the anechoic speech signal.

Although spectral subtraction techniques do not require prior knowledge of the reverberant channel *per se*, an estimate of the reverberation time is necessary for evaluation of the spectral attenuation factor. However, estimation of the reverberation can be very problematic and often cannot be estimated blindly. Spectral subtraction therefore requires a limited amount of prior knowledge about the RIR in order to evaluate the late reverberant energy.

Interestingly, it is noted that the approach in [57] considers the estimation of the reverberant channel from a similar perspective as this dissertation. Although not ex-

plicitly stated in the paper, the RLS channel estimator is of the form of the Kalman correction equations. Chap. 6 derives the optimal channel estimator in the minimum mean-square error (MMSE) sense as a modified version of the Kalman correction equations.

2.4 Source model based blind speech dereverberation

The production mechanism of human speech gives insight into accurate signal modelling as discussed in Chap. 3, allowing for the deduction of prior knowledge about the source signal through an explicit source model. Incorporating explicit speech models in dereverberation algorithms allows for the incorporation of prior information about the speech production mechanism in order to characterise the anechoic signal to be reconstructed.

The spatial enhancement approach by Flanagan [32] discussed in sect. §2.1 utilises a source model to distinguish the speech signal from background noise. Brandstein [58] assumes that the source signal is generated by a filter excited by a glottal pulse train for voiced speech or turbulent noise for unvoiced speech (see sect. §3.3.2 on page 41 for more information on this excitation model for speech). Similar to [32], the model is incorporated in a DSB for spatial enhancement. However, as discussed in sect. §2.1, DSBs can perform poorly in reverberant environments due to the interference of reflections from different directions with the beampath steared at the speech source.

Attias and Deng [59] incorporate an AR speech model pre-trained on a large set of anechoic data in a variational expectation-maximization (EM) framework for single- and multi-microphone speech dereverberation. Experiments indicate that the approach outperforms spatial subtraction in terms of the signal-to-noise ratio (SNR) of the estimated signal compared to the reverberant SNR. However, the approach depends heavily on the training of the model.

Harmonicity is an important structure of speech, particularly with respect to voiced speech components, generated by vibrations of the vocal chords. The frequency formants of clean speech are often approximately multiples of the fundamental frequency [60]. Based on this observation, Nakatani *et al.* [61–65] propose an approach known as harmonicity based dereverberation (HERB) to dereverberate speech by suppressing non-harmonic components in the signal in order to reconstruct harmonic components. By estimating the fundamental frequency, harmonic components are extracted as multiples of f_0 from the distorted speech spectrum for each short-time frame in

the STFT domain. The estimated harmonic speech spectrum is compared to the distorted spectrum in order to compute a dereverberation filter in each time block and frequency bin. By averaging over each time segment, an averaged dereverberation filter is obtained that suppresses reverberation causing non-harmonicity. The inverse acoustic impulse response (AIR) of the reverberant channel is obtained by computing the inverse discrete Fourier transform (DFT) of the dereverberation. By convolving the observed signal with the inverse AIR, an estimate of the direct path signal can be obtained.

HERB applied as a preprocessing step effectively reduces the word error rate of automatic speech recognition (ASR) systems [66]. However, more than 5000 reverberant words – more than 60 minutes of speech – are needed to estimate the dereverberation filter assuming the system is time-invariant [67].

Hopgood [45] observes that identification of the channel and source parameters is possible if the channel and source filter vary at different rates. Consequently, a static infinite impulse response (IIR) channel model for stationary speakers and a block stationary AR (BSAR) source model for speech is assumed. By estimating the AR parameters from the observed signal in each block and plotting the histogram of the estimated parameters with time [68, 69], it can be shown that the channel parameters can be identified by the peaks in the histogram, whilst the BSAR parameters are "smeared" on the floor of the histogram. Hence, it is concluded that the source parameters can be *marginalised* from the joint posterior probability density function (pdf) of the unknown model parameters to obtain the marginal posterior pdf of the channel parameters. The channel parameters are estimated using the Gibbs sampler. The estimated channel parameters are used for inverse filtering the observations in order to obtain a speech signal estimate.

By assuming that the RTF varies slowly as the speaker moves, whilst the source model varies rapidly between blocks, the approach was extended to moving speakers by Evers and Hopgood [1]. As block-stationarity is insufficient to capture the time-varying behaviour of speech, the source model was extended to a block-based time-varying AR (TVAR) model by Evers and Hopgood [1], where the source model varies according to a TVAR process in each block.

It was demonstrated in [1,45,70] that speech from a stationary and moving speaker can be effectively enhanced when distorted by an acoustic horn channel. However, as the approach is based on Gibbs sampling, a batch Markov chain Monte Carlo (MCMC) approach, online processing is not possible. Furthermore, it was shown that approxi-

mately 2000 Monte Carlo iterations were necessary, out of which 10% were considered as the burn-in period.

2.4.1 Issues and insights of dereverberation techniques exploiting explicit source models

A main concern with HERB is the complete elimination of non-harmonic components. Voiced speech naturally contains non-harmonic components. Thus, by eliminating any non-harmonicity from voiced segments, part of the natural sound of speech is removed. Artefacts can thus be perceived in the enhanced signal. Furthermore, due to the assumption of harmonicity, unvoiced speech sounds cannot be approximated by HERB, thus reducing the approach to one type of phoneme only. Another shortcoming of HERB is that fundamental frequency estimation in reverberant environments in itself is problematic as discussed in [71].

Nonetheless, the concept of incorporating harmonic properties of the speech signal for voiced speech only is highly appealing and inspired the development of Chap. 8, where an articulatory-based speech model is used for the modelling voiced phonemes. Unvoiced phonemes are targeted by a Markov-chain based TVAR model in Chap. 7.

The approach exploiting non-stationarity of speech discussed above suffers from its implementation using a rather naïve Gibbs sampler, leading to significant computational burden. Furthermore, the performance of the algorithm decreases in silent periods of the speech signal, i.e., where the channel dictates the over system response. Regardless, the idea of identifying the poles of the channel filter from those of the source filter by exploiting the rapid variation of the source parameters is particularly interesting in scenarios where speakers are either stationary or move reasonably slowly within a room. The development of a more sophisticated and insightful algorithm in Chap. 6 and its extension to moving speakers in Chap. 9 are therefore highly influenced by the concept of exploitation of the non-stationarity in source signals.

2.5 LP residual enhancement

Linear predictive (LP) residual enhancement models the speech production mechanism as an all-pole filter excited by either glottal pulse for voiced speech, or turbulent noise for unvoiced speech (see Chap. 3, and sect. §3.3 on page 38 ff. in particular, for a discussion of all-pole modelling of speech). It is further assumed that the reverberant channel is modelled by an all-zero filter. Therefore, the detrimental effects of reverberation introduce only zeroes to the overall system. Distortion due to additive noise affects the excitation sequence of source filter only. The all-pole filter coefficients are

assumed to be left unchanged as reverberation introduces zeroes rather than poles to the system [72].

As voiced speech is modelled by filtering a glottal pulse train through the all-pole filter, the excitation sequence due to speech is assumed as a well structured pulse train. In contrast, impulses in the distorted signal due to reverberation effects and noise are relatively uncorrelated. Therefore, speech dereverberation can be performed by computing the LP residual of the observed signal and identifying and eliminating the spurious and uncorrelated peaks due to reverberation and noise in voiced speech segments.

Therefore, LP residual techniques estimate the all-pole source filter coefficients using linear prediction analysis. The excitation sequence of the speech signal is thus obtained by inverse-filtering the distorted observed signal with the speech filter. The uncorrelated peaks due to reverberation and noise in the estimated LP residual are attenuated to approximate the clean speech excitation signal. The speech signal can thus be reconstructed by filtering the enhanced excitation signal with the estimated all-pole source filter.

Wavelet transformations can be used to divide a continuous-time signal into different scale components. Each scale component can be assigned a frequency range and can be studied at an appropriate resolution. As wavelets are the derivative of a smoothing function, wavelet transformations describe local extrema of the signal wavelet domain. Impulses in the LP residual of voiced speech can thus be detected as extrema in the wavelet domain. Based on this observation, Griebel [73–76] and Griebel and Brandstein [72,77] propose to use wavelet extrema clustering across multiple channels to obtain a single multi-scale extrema representation.

Rather than processing in the wavelet domain, Yegnanarayana and Sayanarayana [78,79] propose to use the Hilbert transformation for LP reconstruction. The Hilbert transformation is generally used in signal processing to derive an analytic representation of a signal. The resulting Hilbert envelope has large amplitudes at strong excitations in the time-domain signal and can thus be used to detect glottal closure instants (GCIs), i.e., excitation pulses. By applying the Hilbert transformation to a reverberant LP residual, the pulse train structure of voiced speech is thus amplified and the reverberation effects are attenuated, allowing for improved identification of peaks in the residual. This approach is applicable for both single and multiple sensors. As Hilbert envelopes are averaged when utilising multiple microphones, peaks due to reverberation can be further suppressed.

Gillespie [56] use the kurtosis, i.e., the fourth central moment of a distribution, as a measure of the peakiness of the LP residual. As the LP residual of reverberant signal is a time-spread version of the impulse-like LP residual of clean speech, the kurtosis decreases with increasing reverberation. Using an online adaptive gradient descent approach that maximises the LP kurtosis, the reverberation effects can be minimised and the clean speech signal recovered.

Whilst LP residual enhancement in the wavelet domain, by Hilbert envelope weighting, and kurtosis maximisation reduce the effects of reverberation by attenuating the impulses due to reverberation and noise in the LP residual, properties of the underlying structure of speech are not considered in their approaches, such that speech estimates can sound less natural.

Gaubitch and Naylor [30, 36, 80–84] observed that the LP residual between adjacent larynx cycle varies slowly, such that spurious peaks due to reverberation can be temporally smeared by averaging each larynx cycles with its nearest neighbours. Furthermore, suppression of peaks in the *time* domain can be performed by spatial averaging using beamforming techniques. The DSB is thus utilised to partially remove distorting effects from the observed signal. Given the *spatially* averaged signal, the LP residual is computed. Uncorrelated features in the residual are further suppressed by *temporally* averaging the residual over neighbouring larynx cycles. glottal closure instants are extracted from the spatio-temporally averaged residual using the dynamic programming projected phase-slope algorithm (DYPSA).

As LP analysis removes any spurious peaks from the residual, any uncorrelated effects in the clean voiced speech signal are removed as well. Hence, LP analysis suffers from an inherent problem of excessive whitening of the enhanced signal. Delcroix *et al.* thus attempt to circumvent whitening in a multi-sensor approach in [85–89] assume that there is always at least one microphone in a sensor array that is closer to the source than to the noise source. This assumption is valid unless the speaker and noise source are located on opposite sides of the sensor array and are aligned in either x- or y-direction. By obtaining an estimate of the error residual of the source signal, an estimate of the source signal itself is computed. The source parameters can be estimated from the correlation matrix of the output signal. The whitened residual is then applied to the source filter, reintroducing colouration to the signal. For short impulse responses, almost perfect dereverberation can be achieved. However, for longer RIRs, the presence of numerically overlapping zeros among the channels lead to identifiability issues and thus poor dereverberation results [87].

2.5.1 Issues and insights of LP residual enhancement

From a methodological point of view, LP causes non-harmonic parts natural to speech to be ignored and artefacts introduced due to the speech synthesis from the GCIs. Speech estimates can thus sound unnatural. Furthermore, as harmonic signal are targeted by LP residual enhancement, blind dereverberation of *unvoiced* speech, resembling broadband noise, is not solved by this group of algorithms.

From a modelling perspective, the underlying assumption that the LP coefficients are unaffected by reverberation can be ambiguous. It was shown, e.g., by Godsill and Andrieu in [90], that the separation of AR processes (i.e., all-pole filters excited by white Gaussian noise (WGN)) from all-zero mixture models can be performed without ambiguities. However, RIRs are only *approximated* by all-zero models and should, ideally, be modelled using pole-zero models (see sect. §4 for modelling of the RTF). Therefore, realistic channels responses exhibit poles and, hence, modify the poles of the all-pole source model. Therefore, for realistic RIRs, both the source excitation (and hence the LP residual) and the source filter coefficients (and hence the LP coefficients) are modified by reverberation. Synthesis of the source signal using the extracted and unprocessed LP coefficients can therefore never fully suppress the effects of reverberation.

Nonetheless, LP residual enhancement exhibits an interesting parallel to the concept of blind dereverberation proposed in this dissertation. In essence, LP residual enhancement iteratively obtains estimates of the source signal by 1. estimating the source model whilst assuming that the channel is constant and 2. estimating the channel model whilst assuming the source model is constant. In this dissertation, it is proposed that by Rao-Blackwellisation of the joint posterior pdf of the unknown system parameters and signals, the source model parameters and noise terms, the source signal, and the channel parameters can be obtained using three separate estimators. The source model is estimated whilst holding the channel parameters constant. Based on an estimate of the source model parameters, the channel model parameters are evaluated by holding the source model parameters constant. Using the estimates of both the source and channel model parameters, the source signal is obtained.

An even stronger analogy exists between LP residual enhancement and the blind dereverberation approach exploiting non-stationarity in [1] and discussed in sect. §2.4. In this approach, the source model and channel model are iteratively estimated by 1. estimating the source model by assuming the channel as a nuisance parameter, and 2. estimating the channel model by assuming the source as a nuisance parameter. Therefore, strong analogies exist between the dereverberation approach proposed

in this dissertation and LP residual enhancement.

2.6 Conclusions

Dereverberation approaches in the literature can be classified by the following groups:

1. spatial filtering, using beamformers to attenuate signal reflections arriving from different directions in order enhance the direct-path signal (e.g., [34,35]),
2. homomorphic transformations, reflecting the non-additive mixture of speech with a reverberant channel to a domain where the mixture can be considered as additive (e.g., [37]),
3. spectral enhancement, modifying the short-term spectrum of reverberant speech in order to remove late reverberation effects (e.g., [50,67]),
4. explicit source modelling, exploiting properties of the source production mechanism to identify the anechoic speech signal from the reverberant effects (e.g., [1,64]), and
5. linear prediction residual enhancement, distinguishing peaks due to the excitation of the vocal tract from spurious peaks due to reverberation (e.g., [72,83,86]).

As reverberant reflections arrive from multiple positions in the room and are likely to interfere with the beam path, beamforming approaches only remove reverberation to some degree. Nonetheless, the concept of accumulating statistical evidence of the same statistical event by using multiple sensors inspires the use of microphone arrays for improvement of the dereverberation performance in Chaps. 7 and 8. In order to evaluate the computational burden several microphones incur on the proposed algorithm, Chap. 10 investigates the computational complexity of the proposed algorithm for a variable number of sensors. Inspired by the concept of spectral enhancement, a sub-band implementation of the dereverberation framework is proposed in order to reduce the computational complexity.

As an approach to direct dereverberation, spectral enhancement based on spectral subtraction provide the ability to retain early reflections in the estimated signal, whilst eliminating late reverberation only. However, partial knowledge of the RIR is required, which is generally not available in practice.

Dereverberation by explicit source modelling often suffers from either the necessity of excessive training or the restriction to one model only. The idea of incorporating prior information about the speech production mechanism is, however, highly attractive and therefore influences the development of the proposed dereverberation algorithm by explicitly modelling unvoiced and voiced speech in Chaps. 7 and 8. Furthermore, the consideration that non-stationarity of the vocal tract can be exploited for identifiability between the source and channel parameters is a fundamental assump-

tion of the proposed algorithm in Chap. 6.

LP residual enhancement techniques are often based on *synthesising* the speech signal from the identified excitation peaks, thus neglecting natural speech components in the synthetically generated source signal estimate. Furthermore, as the main underlying assumption is that the reverberant channel consists of zeros only and therefore does not influence the coefficients of the source model, LP residual techniques face identifiability issues of resonant peaks in realistic RIRs. As spurious peaks due to reverberation can only be detected in harmonic data streams, LP residual enhancement is also unsuitable for unvoiced speech.

That having said, the concept in LP residual enhancement of estimating the source model whilst holding the channel constant and *vice versa* has a strong parallel to the proposed algorithm, where the source model, channel model, and source signal are estimated using three different estimators, each of which assumes that an estimate of the remaining two variables is known.

Many of the existing approaches attempt to estimate the RTF from the observed signal in order obtain a source signal estimate by inverse filtering of the observed signal with the reverberant channel estimate. However, amongst other issues, inverse filtering inherently leads to scaling of errors in channel estimate, potentially increasing distortion in the enhanced signal. It is therefore highly desirable to estimate the source signal directly as opposed to its reconstruction by inverse filtering with a channel estimate. The proposed dereverberation approach in Chap. 6 is therefore based on direct and optimal estimation of the source signal using the Kalman filter.

The take-home message of this chapter is two-fold: On the one hand, several highly appealing ideas are taken away and incorporated from a different perspective in the dereverberation approach proposed in this dissertation. On the other hand, it was shown that most dereverberation algorithms in the literature suffer from very restricting underlying assumptions that theoretically and practically preclude extensions to different scenarios, e.g., to different phoneme types, time-varying source-sensor positions, or either single- or multi-sensor processing. Furthermore, extremely few of the discussed algorithms facilitate real-time processing, crucial in, e.g., military and security applications.

As already elaborated on in Chap. 1, the underlying aim of this thesis is therefore to investigate a flexible dereverberation framework allowing for extensions to different scenarios, whilst embodying appealing properties of the approaches in the literature.

Part II

Models and methodology

Chapter 3

Speech production, signals, and models

3.1 Introduction

Human speech is a highly dynamic process, involving various information transfer stages. On a linguistic level, the message is formed in the speaker's brain in a discrete or symbolic form [91]. The corresponding instructions are transferred to the articulators – i.e., lips, tongue, larynx, jaw, etc. – and translated from discrete to continuous movements on a physiological level. As a result, air is pushed from the lungs through the vocal tract, forming sound by the time-varying articulators, and transmitted through a distorting noisy and, or, reverberant channel on an acoustic level. Finally, on the audiological and perceptual level, the sound produced on the acoustic level is received at the listener's and speaker's ears where it is used for feedback control. The sound is translated to mechanical motion by the ossicles of the middle ear, to fluid pressure waves in the medium bathing the basilar membrane of the inner ear and invoking travelling waves. The travelling waves stimulate hair and hence trigger electrical, mechanical, and biochemical reactions of the auditory nervous fibres. The neural responses are finally used at higher processing stages in the brain [91].

In order to make sense of and interpret these stages, mathematical models are used as a characterisation and abstraction of each stage and can be used for speech analysis, enhancement, recognition, synthesis, coding, production, and perception. Detailed discussions, especially in the field of speech recognition, can be found in, for example [92–98] in chronological order. In these references and in general speech is commonly either modelled in terms of the physiological detail of the speech *production* (or articulatory) system as discussed in sect. §3.2, or by modelling the speech *signal*, as discussed in sect. §3.3.

Speech system models offer high spectral fidelity as formants are shaped accu-

rately according to physical reality, proving particularly useful for speech synthesis. However, without a linguistic layer in the physical model, articulatory events are often difficult to predict and model, such that speech production systems can prove inefficient for modelling of distorted speech signals in estimation frameworks. Instead of modelling the speech production mechanism, speech signal models represent formants by digital filters used for approximation of uttered resonances. Depending on the sophistication of the filter, anomalies in the utterances might not necessarily be captured. Nonetheless, speech signal models can provide analytical tractability, which is crucial to many Bayesian estimation approaches and are therefore utilised in Chaps. 6 and 7 for construction of the reverberant speech system model. The concepts discussed in sect. §3.2 are utilised in Chap. 8, where the model developed in Chaps. 6 and 7 is extended to a model incorporating prior physical knowledge about the human speech production mechanism in the signal model.

3.2 Speech production models

Speech production models describe the movement of the human articulators in terms of pressure waves, particle- and volume-velocities. As the articulators are described in physical terms, co-articulation effects that are difficult to formulate in a mathematical framework, occur naturally in accurate speech production models. Speech production models are hence particularly useful for capturing the dynamic properties of speech.

Each language consists of a finite number of speech sounds unique to the speaker but distinguishable for a listener familiar with the language. The most basic linguistic elements are called *phonemes* [99]. The concatenation of phonemes to generate words and sentences is linguistically organised and structured by the central nervous system using acoustic feedback of the hearing apparatus and the speech musculature [100]. For the production of phonemes, the respiratory apparatus acts as a motor of the vocal tract apparatus. The vocal tract apparatus, physically structuring airflow from the respiratory apparatus, consists of *phonatory* and *articulatory* organs and can be broadly divided into three regions:

- the sub-glottal tract, i.e., the lower respiratory tract below the glottis that includes the lungs, trachea, and bronchial tubes;
- the glottal tract from the glottis to the lips; and
- the nasal cavities.

Airflow, which can be considered as an unstructured source signal, is produced in the sub-glottal tract in the lungs, travelling from the glottis to the trachea. The cavities between the glottis and lips form a complex three-dimensional tube, consisting of immobile walls such as the dental arch and palatal dome, rigid walls allowing for subtle

changes, such as the pharyngeal wall, as well as soft walls such as the tongue, velum, uvula, lateral pharyngeal wall, and lip tube. Depending on the changes in the glottal tract, different speech sounds are generated:

- Voiced sounds are produced by airflow causing vibrations of the vocal cords, thus modulating the stream of air into discrete puffs or pulses;
- Unvoiced sounds are produced by air forced through a constriction in the vocal tract, producing turbulent flow and incoherent sounds;
- Plosive sounds are produced by the abrupt release of pressure behind a complete closure in the articulatory system.

The width of the jaws relative to the pharyngeal cavity affects tongue articulation and variation of the vocal tract shape. The mobility of the jaw relative to the skull varies openness of vowels. Furthermore, the size of the tongue relative to the oral cavities determines the articulatory space for vowels. From a modelling perspective, the glottal tract can therefore be considered as an acoustic filter that structures the unstructured airflow from the sub-glottal tract. The nasal cavity acts as an "accessory channel" [101] to the glottal tract, building additional resonances to produce nasal sounds.

The human speech apparatus is therefore a complex system, consisting of organs of *phonation*, i.e., voice production, and organs of *articulation*, i.e., settings of the speech organs. The sound excitation from the lungs is emitted into the vocal tract resembling a tube. In its most general form, the vocal tract can be described by the propagation of waves in a flared horn. Using Newton's laws, the pressure can thus be derived in terms of a partial differential equation (PDE) (see Appendix A.2), often referred to as the Webster equation [100, 102]:

$$\frac{A(x)}{c^2}\frac{\partial^2 p(x,t)}{\partial t^2} = \frac{\partial}{\partial x}\left[A(x)\frac{\partial p(x,t)}{\partial x}\right] \tag{3.1}$$

where $p(x,t)$ is the sound pressure dependent on the distance, x, and time, t, $\mathbf{v}$ is the vector velocity of an air particle, and $A(x)$ is the vocal tract area as a function of the distance is the sound pressure. Note that in the derivation of eqn. (3.1), transverse modes are ignored. The reasoning behind this assumption is as follows: If the wavelength of sound is large compared to the diameter of the tube, the propagation of sound can be modelled as a planar propagation in just one dimension, such that transverse modes can be ignored. The length of the vocal tract from the glottis to the lips lies between $14-15$cm in female adults and between $16.5-17.5$cm in male adults [103]. Assuming a subglottal resonance frequency of $f_R = 2100$Hz [104], at a body temperature of 35°C the speed of sound in air at sea level is $c = (331 + 0.6 \times 35°\text{C})\,\text{m/s} = 352\,\text{m/s}$. Therefore, the wavelength, λ, of sound in air corresponds to $\lambda = c/f_R = 16.76$cm. Thus, the wavelength of sound in the vocal tract is significantly larger than the vocal tract diameter of

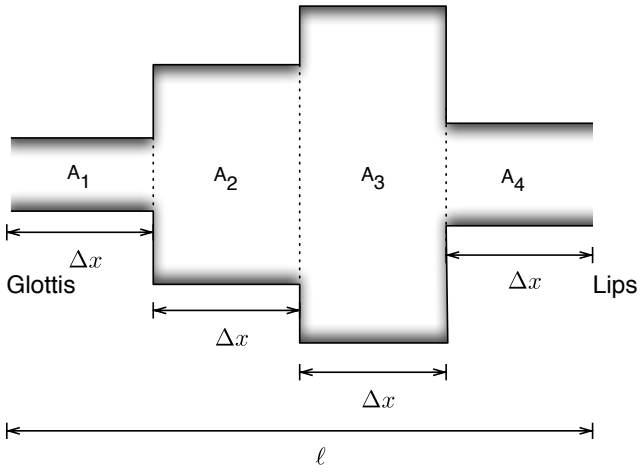


Figure 3.1: *Concatenation of lossless acoustic tubes of equal lengths, Δx [107].*

about $2-2^3/_4$ cm [105]. Wave propagation in the vocal tract can thus be approximated by planar propagation when neglecting parts of the tract with large width [100, 106].

Nonetheless, Webster's equation can only be solved if $A(x)$ is a well-behaved analytic function. The properties of the vocal tract are changing continuously such that the area of the acoustic tube cannot be expressed in closed form. Therefore, numerical approximation is necessary. One of the most widely used speech production models assumes that the continuous area of the vocal tract can be segmented into small uniform sections as depicted in Fig. 3.1.

3.2.1 Lossless acoustic tube model

A simplified model of the human vocal tract [107, ch. 3.3] assumes that the vocal tract can be represented by a *concatenation* of K lossless acoustic tubes of constant cross-sectional areas, $A(x) = A_k$, $k \in \mathcal{K}$ as illustrated in Fig. 3.1. For sufficiently many segments the concatenation of the constant cross-sectional areas approximates a slowly time-varying circular acoustic tube that can be used to model the human vocal tract as proposed by Kelly and Lochbaum [108]. The model can be modified so that losses due to friction, heat conduction or wall vibration, influencing the bandwidths of the modelled resonances can be accounted for at the glottis and lips [107, ch. 8].

3.2.1.1 Relevance to thesis through reflection coefficients

Having a closer look at the sound waves at boundaries between tube sections, part of a travelling wave propagates through to the next tube section as a wave front meets the discontinuity area of a section. The remainder is reflected back into its own sec-

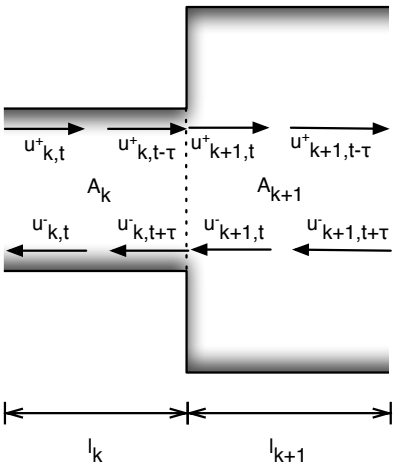


Figure 3.2: *Sound propagation in concatenated lossless acoustic tubes*

tion. A wave only propagates fully if the impedance of the next section meets that of the previous section, i.e., the cross-sectional areas $A_k = A_{k+1}$ [109]. Fig. 3.2 illustrates how the flow of volume velocity, $u^+_{k,t-\tau}$, $\tau \geq 0$ in section $k \in \mathcal{K}$ partial traverses into section $k+1$ of different cross-sectional area, whilst part of the wave is reflected into volume velocity, $u^-_{k,t-\tau}$ traveling in the opposite direction.

The idea of reflections between tube segments of different cross-sectional areas leads to the concept of the so-called reflection coefficient between two sections, indicating the ratio between wave propagation and reflection. The volume velocity, $u_{k,t}$, can be expressed as a PDE of similar form to the pressure velocity in Webster's equation in eqn. (3.1). Solving the resulting PDE for $u_{k,t}$, and comparing with the volume velocity in the right adjacent tube segment, $u_{k+1,t}$, the reflection coefficient, r_k, can be derived as illustrated in Appendix A.3.1 and is found as:

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{3.2}$$

As cross-sectional areas are positive by definition, the right hand side (RHS) in eqn. (3.2) takes its maximum at $+1$ as $A_{k+1} \gg A_k$ and its minimum at -1 as $A_k \ll A_{k+1}$. Therefore, the reflection coefficient is a real number satisfying $-1 \leq r_k \leq 1$. If $A_k = A_{k+1}$, the reflection coefficient $r_k = 0$ the travelling wave is completely transmitted [110].

Reflection coefficients are used in physics and engineering to describe wave propagation in media with discontinuities. For instance, in telecommunications, reflection coefficients are utilised for transmission lines in order to describe the ratio of

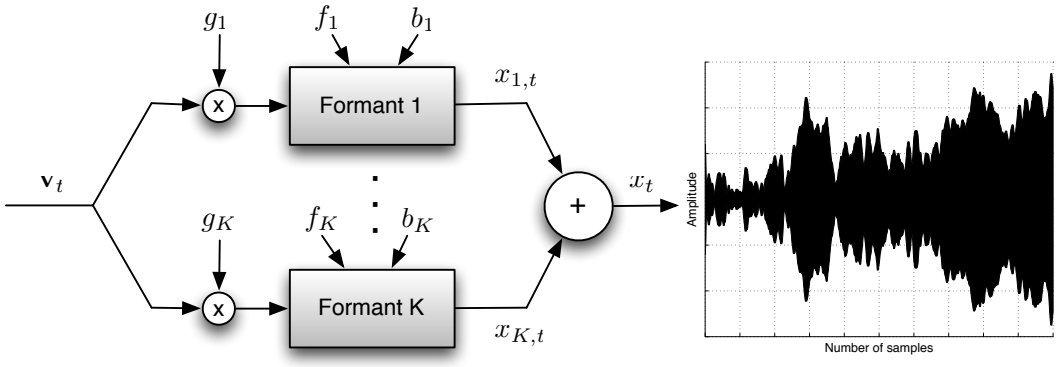


Figure 3.3: *Parallel formant synthesiser [111]*

impedance between the source and the load [107]. In Chap. 8 of this thesis reflection coefficients play a vital role in that a partial correlation (PARCOR) speech model, generally expressed in terms of reflection coefficients, is utilised. PARCORs modles are phrased in terms of infinite impulse response (IIR) lattice filters. The lattices are connected by reflection coefficients, which are identical to the reflection coefficients of the human vocal tract in eqn. (3.2).

3.2.2 Parallel formant synthesiser model

As opposed to the acoustic tube model in sect. §3.2.1 where the vocal tract was modelled, formant synthesisers model the formants, i.e., the spectral peaks, generated in the vocal tract. The three to five formants in the speech spectrum are modelled by means of three to five resonant circuits with variable frequency and amplitude. Therefore, the vocal tract transfer function is simulated by a sequence of second-order filters. For cascaded connections between filters, the transfer function of formant synthesisers resembles that of the vocal tract without nasal coupling [102, 112]. Cascaded formant synthesisers automatically control the formant amplitudes by adjusting the bandwidths appropriately. However, due to the omission of of nasal coupling, stop and fricative sounds are not modelled accurately and require the introduction of additional resonators to introduce extra poles and zeros in the transfer function (see sect. §3.3.3 for more information on the relation between poles, zeros, and nasal sounds). Parallel synthesisers (see Fig. 3.3) are preceded by an amplitude control, specifying the relative amplitude of the spectral peak (or formant) [113]. Additional resonators for modelling nasals, stops, or fricatives are therefore not necessary. However, the formant frequencies need to be explicitly specified.

In order to accurately specify the formant frequencies, typical values in human speech are consulted. The fundamental and formant frequencies are visible in the

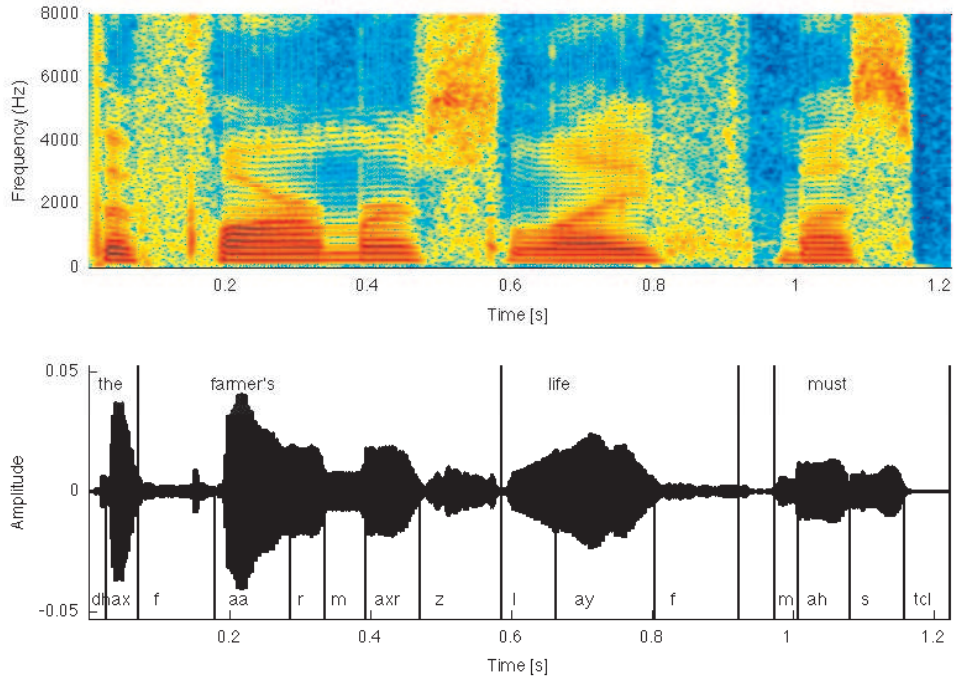


Figure 3.4: *Spectrogram (top) and time-domain signal (bottom) of 'The farmer's life must' uttered by a female at* $f_s = 16kHz$*; Red areas: high energy, blue areas: low energy.*

short-term spectrum of speech as spectral lines. According to [60], vowels generally exhibit fundamental frequencies, f_0, between $136-141$Hz in males and $210-235$Hz in females. The first formant frequency, f_1, lies between $270-730$Hz in adult males and $310-860$Hz in adult females. f_2 lies between $840-2290$Hz in males whilst females exhibit a f_2 of $920-2790$Hz. The f_3 range for male vowels is between $1690-3010$Hz and $1960-3310$Hz for females.

In most cases, most energy is concentrated first two formant frequencies, f_1 and f_2, such that f_1 and f_2 are sufficient to disambiguate vowels. Therefore, special attention must be paid in formant synthesisers to model f_0-f_2 accurately. For illustration, Fig. 3.4 shows the spectrogram of the sentence "The farmer's life must" uttered by a female American speaker at 16kHz sampling rate. Three to five formant tracks are visible for voiced phonemes [102] such as /ə/ in "the". The short-time spectrum of the speech signal of, e.g., /aː/ in "farmer" and /ʌɪ/ in "life" reveals that most energy lies between approximately $200-2000$Hz, i.e., f_0-f_2, with little energy in f_3 and f_4.

Formant synthesisers are mostly used for text-to-speech interfaces. Particularly well known for their real-time speech synthesis using formant synthesisers are the Sega arcade systems developed in the 1980s. Although improved speech *synthesisers* have been developed since, e.g., using hidden Markov model (HMM)-based synthesis,

formant synthesisers are notably accurate at *modelling* human speech from a physical perspective. parallel formant synthesizers (PFSs) are utilised in this thesis in Chap. 8, where the speech signal model utilised in Chaps. 6 and 7 is extended to a more physically meaningful model, exploiting prior knowledge available about the vocal tract.

Overall, a physical description of the articulators allows for modelling of the human vocal tract for speech synthesis. Nonetheless, as the source signal is generated as a side-product and actually not directly described by the model, speech production models are difficult to apply directly for speech signal estimation, such as the dereverberation problem in this thesis. For these applications, it is therefore desirable to model the source signal directly, rather than the speech production mechanism.

3.3 Speech signal models

Signal models can generally be divided in two classes: parametric and non-parametric models. Parametric models model the signal as the output of autoregressive moving average (ARMA) models (or variants thereof). The properties of the filter and hence characteristics of the speech signal are determined by a finite set of parameters associated with the model [114]. The parameters are chosen to reflect prior knowledge available about the system. If no prior knowledge is available, the choice of any underlying pre-specified mathematical model might be unjustified. In such cases, *non-parametric* models should used instead as no assumptions on the underlying structures of the data are assumed. Histograms [115] or higher order statistics [116–118] are examples of statistics considered in non-parametric models.

For speech signals, the speech production mechanism and resulting signal properties are well researched and prior information about the speech production system can be assumed (see sect. §3.2 and references therein). Parametric models are therefore well suited for modelling speech signals. It is therefore desirable to rephrase the physical model of the human vocal tract as a concatenation of lossless acoustic tubes in sect. §3.2 in terms of a parametric model. Using the transfer function of the acoustic tube model, the speech signal can be expressed as a linear combination of past signal samples with a set of source parameters specifying the properties of the signal.

3.3.1 Transfer function of the vocal tract

Any transfer function is described as the ratio of the input and output of the system. Recalling that the human vocal tract can be interpreted as a concatenation of K acoustic tubes of equal lengths as shown in Fig. 3.1 on page 34, the input of the system is the airflow generated in the glottis, whereas the output is the signal at the lips. There-

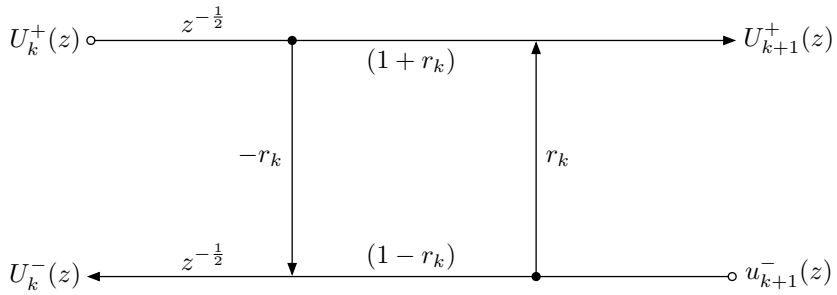


Figure 3.5: *Flow chart illustrating relationship between z-transforms at a junction between two segments of the acoustic tube after [107].*

fore, the transfer function of the vocal tract, $V(z)$, can be expressed as the ratio of the response at the lips, $U_L(z)$, and the response at the glottis, $U_G(z)$, i.e.,

$$V(z) = \frac{U_G(z)}{U_L(z)}. \tag{3.3}$$

As already explained in sect. §3.2, in particular using Fig. 3.2, at each junction between tubes an acoustic wave partially propagates through to the adjacent tube segment, whilst part of it reflected back according to the reflection coefficient, r_k. Therefore, at a random point in the acoustic tube $k+1$, where $k \in \mathcal{K}$, the response of tube k travelling from the glottis to the lips, $U_k^+(z)$, as well as the response reflected at the junction to the adjacent tube, $U_{k+1}^-(z)$, are observed. The derivation of the responses in terms of the reflection is beyond the scope of this thesis, however, a detailed derivation can be found in [107, ch. 3]. Accordingly, the propagated and reflected transfer functions of the responses, $U_{k+1}^+(z)$, and $U_k^-(z)$ respectively, can thus be expressed as:

$$U_{k+1}^+(z) = (1+r_k)\, z^{-\frac{1}{2}}\, U_k^+(z) + r_k\, U_{k+1}^-(z) \tag{3.4a}$$

$$U_k^-(z) = -r_k\, z^{-1}\, U_k^+(z) + (1-r_k)\, z^{-\frac{1}{2}}\, U_{k+1}^-(z). \tag{3.4b}$$

This relationship is shown graphically in Fig. 3.5, illustrating a lattice resembling structure between the propagated and reflected responses. As one seeks the transfer function of the system in eqn. (3.3), the response of the glottis, $U_G(z)$, and lips, $U_L(z)$, should therefore be expressed in terms of eqn. (3.4).

By rephrasing eqn. (3.4) in terms of matrix description, expressions for $U_G(z)$ and $U_L(z)$ can be derived as shown in Appendix A.3.2. The transfer function, $V(z)$, can

therefore be expressed as [107]:

$$V(z) = \frac{\left\{\frac{1}{2}(1+r_G)\prod_{k=1}^{K}(1+r_k)\right\} z^{-K/2}}{D(z)} \tag{3.5}$$

where the concatenated tubes are simplified to sections of equal lengths, $\Delta x = \ell/K$. The denominator is defined as

$$D(z) = \underbrace{\begin{bmatrix}1\\-r_G\end{bmatrix}^T \underbrace{\begin{bmatrix}1 & -r_1\\ -r_1 z^{-1} & z^{-1}\end{bmatrix}}_{\mathbf{P}_1(z)} \cdots \begin{bmatrix}1 & -r_{K+1}\\ -r_K z^{-1} & z^{-1}\end{bmatrix}}_{\mathbf{P}_K(z)} \begin{bmatrix}1\\0\end{bmatrix} \tag{3.6}$$

where r_G denotes the reflection coefficient in the glottis, r_1 is the reflection coefficient of the tube immediately following the glottis, r_{K+1} is the reflection coefficient of the last tube segment, and $\mathbf{P}_k(z) = \begin{bmatrix}D_k(z) & -z^{-k}D_k(z^{-1})\end{bmatrix}$ with $D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1})$. Hence, at the last section:

$$D(z) = \mathbf{P}_K(z)\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}D_K(z) & -z^{-K}D_K(z^{-1})\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} = D_K(z) \tag{3.7}$$

Assuming that $r_G = 1$ at the glottis, it can be shown by manipulation of the matrices in eqn. (3.6) that the denominator of the acoustic tube model can be expressed in terms of the following recursions (see Appendix A.4.4):

$$D(z) = D_K(z) \tag{3.8a}$$

$$D_0(z) = 1 \tag{3.8b}$$

$$D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}) \qquad k = 1, \ldots, K \tag{3.8c}$$

which is equivalent to the autoregression:

$$D(z) = 1 - \sum_{k=1}^{K} \alpha_k z^{-k}, \tag{3.9}$$

where α_k are combinations of the reflection coefficients, r_k. Considering, for example, the case where $K = 2$, then $\alpha_1 = -r_1(1 + r_2)$ and $\alpha_2 = -r_2$. By approximating the vocal tract by a concatenation of lossless acoustic tubes of equal lengths, the transfer

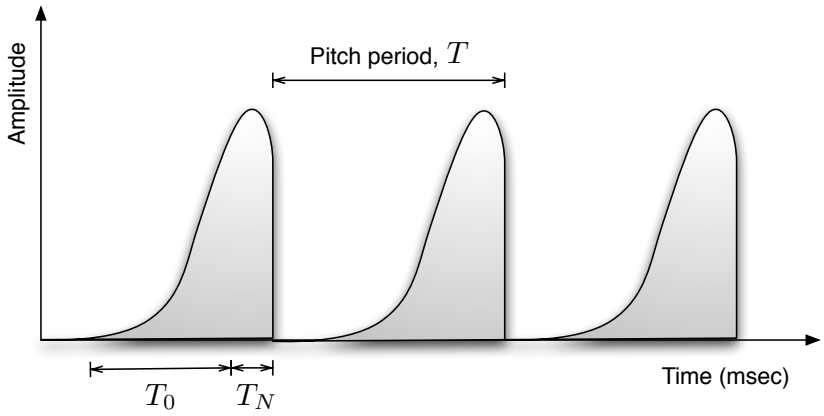


Figure 3.6: *Glottal pulse waveform; Pitch period,* T*; Interval between open and closed phase of the vocal folds,* T_0*; Period between peak and end of the open phase,* T_N.

function of the vocal tract, $V(z)$, can be expressed as a rational function, i.e.,

$$V(z) = \frac{\left\{\frac{1}{2}(1+r_G)\prod_{k=1}^{K}(1+r_k)\right\} z^{-K/2}}{1-\sum_{k=1}^{K}\alpha_k z^{-k}}. \tag{3.10}$$

Hence, the transfer function of the vocal tract is of the form of an all-pole filter [107] and the source signal corresponds in time-domain to an autoregressive (AR) process.

3.3.2 AR representation of speech

The transfer function of the tube model therefore corresponds to an all-pole filter with delays corresponding to the number of sections of the model. Thus, by simplifying the lossless acoustic tube model of the human vocal tract to sections of each length, AR processes can be used to represent the transfer function of the model. Local correlations in the signal are exploited by modelling the current signal sample as a linear combination of the previous samples [119]. AR models are popular models in the speech processing community as they accurately capture the short-term spectrum of speech (see, e.g., [107, 120, 121] and references therein). The transfer function in eqn. (3.5) can therefore be generalised to

$$V(z) = \frac{\gamma z^{-\frac{K}{2}}}{1-\sum_{k\in\mathcal{K}}\alpha_k z^{-k}}. \tag{3.11}$$

The model excitation varies with the voicing of the speech signal. As discussed in sect. §3.2, human speech can be broadly divided into voiced sounds, generated by airflow causing vibrations of the vocal cords, and unvoiced sounds, produced by air

forced through a constriction in the vocal tract and resembling turbulent flow. The airflow generated at the glottis resembles the glottal waveform in Fig. 3.6: as the vocal cords open, the airflow increases slowly. Closely after reaching its maximum amplitude at full opening, the folds close and the waveform plummets to zero amplitude. The glottal waveform can therefore be described as (see [122] or [123, ch. 3.2.2]):

$$g_t = \begin{cases} 3\left(\frac{t}{T_0}\right)^2 - 2\left(\frac{t}{T_0}\right)^3 & \text{if } 0 \le t \le T_0 \\ 1 - \left(\frac{t-T_0}{T_N}\right)^2 & \text{if } T_0 < t \le T_0 + T_N \end{cases} \tag{3.12}$$

where g_t denotes the glottal waveform, T_0 is the pitch period, i.e., the duration of one glottal cycle of open and closed phase, and T_N is the time interval between the peak and the end of the open phase.

During voiced periods, speech can thus be modelled by exciting the acoustic tube model of the vocal tract by the periodic glottal pulse waveform in eqn. (3.12). During unvoiced periods, the excitation signal is turbulent noise. Typically, a random sequence with flat spectrum, such as white Gaussian noise (WGN), is used. In order to impose time-varying properties of the unvoiced source, the variance of the WGN can be modelled as a first-order Markov chain, varying depending on the value of the standard deviation at the previous time step and driven by WGN. By definition, variance terms are bound between $0 \le \sigma^2 \le \infty$. In order to reinforce this constraint, the unvoiced excitation should be sampled from the log-variance rather than the variance directly. Therefore, the log-variance of the source excitation for unvoiced speech can be modelled as:

$$\phi_{v_t} = \phi_{v_{t-1}} + \sigma_{\phi_{v_t}} r_{\phi_{v_t}}, \qquad r_{\phi_{v_t}} \sim \mathcal{N}(0, 1) \tag{3.13}$$

or, equivalently in form of a pdf:

$$p\left(\phi_{v_t} \mid \phi_{v_{t-1}}\right) = \mathcal{N}\left(\phi_{v_t} \,\middle|\, \phi_{v_{t-1}}, \sigma^2_{\phi_{v_t}}\right) \tag{3.14}$$

where $\phi_{v_t} \triangleq \ln \sigma^2_{v_t}$ is the logarithmic value of the excitation variance, $\sigma^2_{v_t}$, and $\sigma_{\phi_{v_t}}$ is assumed constant and known. The initial state of the chain is assumed as $p(\phi_{v_0}) = \mathcal{N}\left(\phi_{v_0} \,\middle|\, 0, \sigma^2_{v_0}\right)$. The variance, $\sigma^2_{v_t}$, of the source excitation, v_t, can thus be obtained by drawing random samples from the log-variance in eqn. (3.13) and transforming to linear domain via $\sigma^2_{v_t} = \exp\{\phi_{v_t}\}$.

The alternation between voiced and unvoiced states (or open and closed phases) can be incorporated in the AR speech model as a switch between a glottal pulse gener-

ator, with response $G(z)$ corresponding to the z-transform of eqn. (3.12) and a random noise generator with response $N(z)$ [106]. The response at the output of the lips can therefore be expressed as

$$X(z) = \begin{cases} G(z)\,V(z) & \text{if voiced,} \\ N(z)\,V(z) & \text{if unvoiced} \end{cases} \tag{3.15}$$

Note that a filter of the form $R(z) = 1 - z^{-1}$ can be incorporated in this response to account for the radiation model of the lips [123]. The speech production model incorporating the switch between voiced and unvoiced excitation is illustrated in Fig. 3.7.

In time domain, the speech signal corresponds to the difference equation:

$$x_t = \sum_{q \in \mathcal{Q}} a_q\, x_{t-q} + \sigma_{v_t}\, v_t, \tag{3.16}$$

for $t > Q$, where the initial conditions can be set to $x_t = 0$ for $t \leq Q$, and where $\mathbf{x}_{0:t} = \begin{bmatrix} x_0 & \dots & x_t \end{bmatrix}$ is the trajectory of speech signal samples for $t \geq 0$, $\{a_q\}_{q \in \mathcal{Q}}$ are the AR parameters of the model of order Q, and $\sigma_{v_t}\, v_t$ is the model excitation from the glottis with standard deviation σ_{v_t}. v_t either takes the form of the glottal waveform in eqn. (3.12) for voiced periods, or WGN for unvoiced periods, i.e., $v_t \sim \mathcal{N}(0, 1)$. The model parameters can be specified to enforce certain desired properties or even dynamics on the model and will be discussed in more detail from sect. §3.3.4 onwards.

The general all-pole model developed thus far forms the basis for a majority of speech recognition, analysis and synthesis systems to date. The popularity of the AR speech model is not at least due to its extendibility, e.g., to a multitude of vocal tract models specified by the AR parameters. From a mathematical perspective, the linearity of the signal (see eqn. (3.16)) allows for mathematically tractability. This is of particular interest in this thesis as the algorithmic framework relies on mathematical expressions of the underlying speech production and reverberation systems. Nonetheless, it should be noted at this point that the linearity of the speech model stems from the underlying assumption that the excitation source is independent of the vocal tract system. In reality, coupling effects due to aeroacoustic events between the glottal source and vocal tract occur, leading to a far more complex and, in fact, nonlinear relationship (see, e.g., [123–125]). In general, the model nonetheless provides a sufficiently accurate representation of the human vocal tract, such that these effects are usually disregarded. Another aspect of coupling, in this case between the nasal and oral cavities, is sometimes resolved by the introduction of zeros in the speech model.

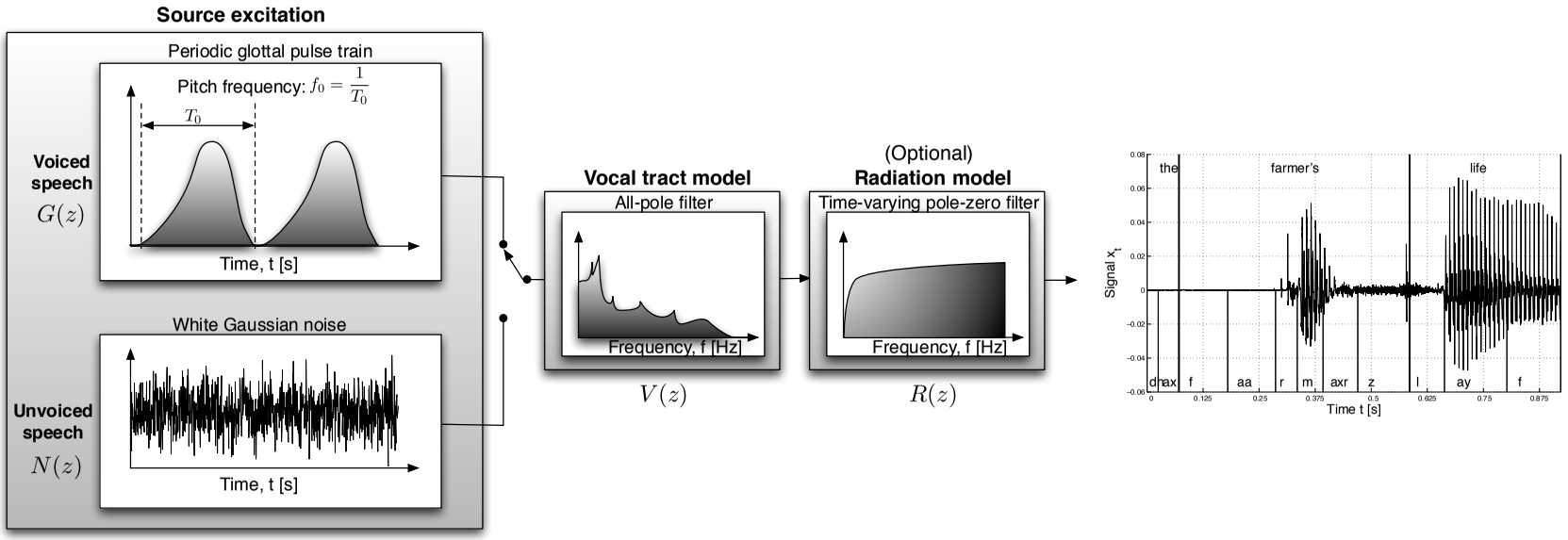


Figure 3.7: *Speech production model of voiced and unvoiced speech, switching between white Gaussian noise excitation and glottal pulse train as excitation sequence after [123].*

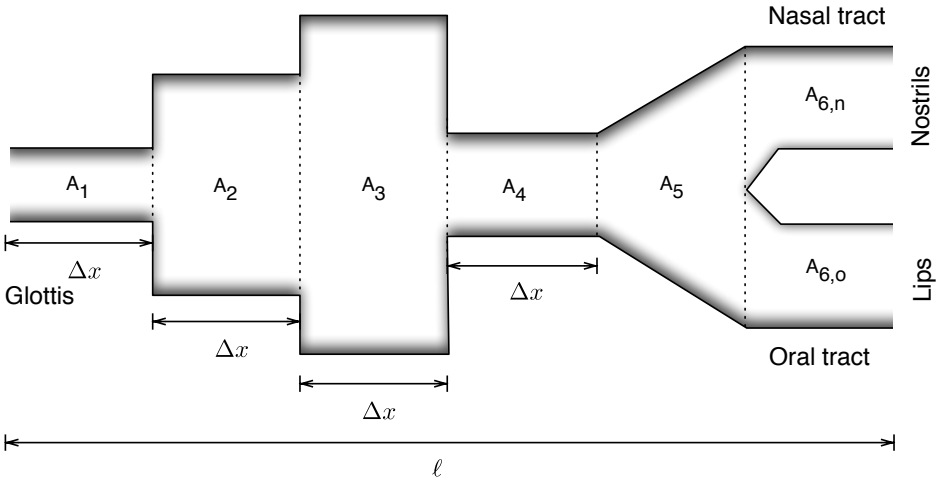


Figure 3.8: *Lossless acoustic tube model including nasal tract*

3.3.3 ARMA speech model

The speech model developed in sect. §3.3.2 contains poles only, but has no zeros. Poles define the resonances or formants of the model, whereas zeros describe the antiresonances or troughs of the transfer response. Antiresonances are introduced in the transfer function when side chambers open in the acoustic path. Sound is thus absorbed near the antiresonant frequencies (particularly at higher frequencies), decreasing the spectral energy and hence the amplitude of sound. Thus far, the vocal tract is modelled as one closed cavity, excited from the glottis and radiating sound through the oral tract from the lips. For voiced pronunciation, the velum is thus assumed closed and sound is radiated through the oral tract only.

For nasal sounds, the velum does not completely close the pharyngeal passage to the nasal tract, thus allowing the nasal cavity to act as a side chamber to the oral cavity resonator. As a side chamber is opened, antiresonances will occur. However, due to the negligence of zeros in the model, these antiresonances cannot be model using an AR process, leading to a model mismatch for nasal sounds. Although many languages, such as English, are mainly based on non-nasal sounds, the inability to model nasal sounds can be problematic particularly in French [126].

Nasal sound production can be incorporated in the model by introducing zeros to the transfer function. Effectively, the introduction of zeros partitions the last tube segment before the lips into two tracts, i.e., the nasal and oral tract as shown Fig. 3.8.

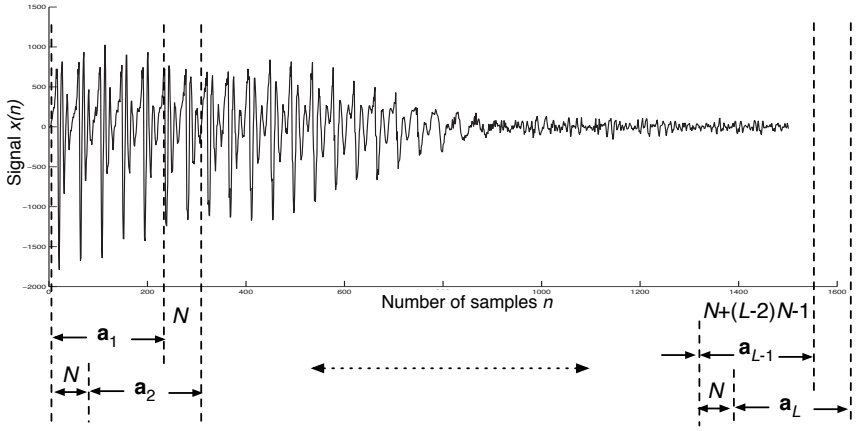


Figure 3.9: *Estimation of time-varying parameters of source signal using stationary AR modeling where* L *is the number of blocks,* N *is the step size between windows, and* $\mathbf{a}_i$ *are the* Q *source coefficients in blocks* $i = 1, \ldots, L$.

The corresponding ARMA process is given by:

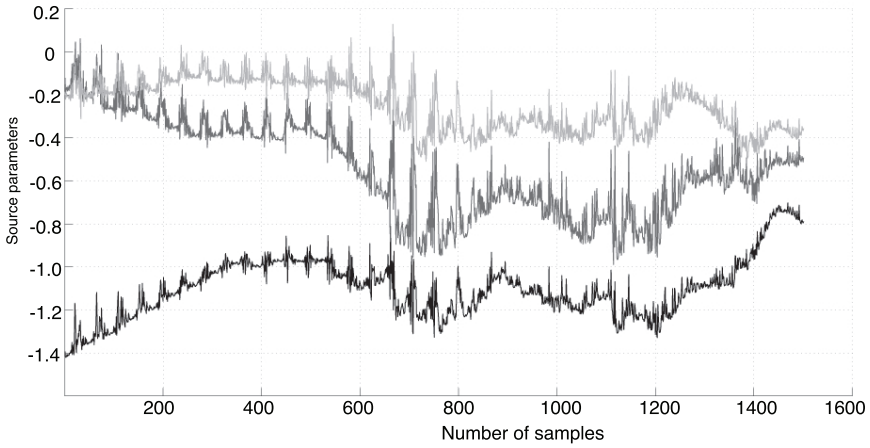
$$x_t = \sum_{q \in \mathcal{Q}} a_q \, x_{t-q} + \sum_{r \in \mathcal{R}} d_r \, \sigma_{v_{t-r}} \, v_{t-r} \tag{3.17}$$

Nonetheless, the inclusion of zeros in the model can lead to non-uniqueness ambiguities in the pole-zero pairs. Despite the exclusion of nasal sounds, speech models therefore often resort to the AR model in eqn. (3.16). The results presented in this thesis are based on non-nasal utterances from native English or American speakers. In order to avoid identifiability issues due to the inclusion of zeros in the model, the remainder of this thesis therefore focuses on AR models.

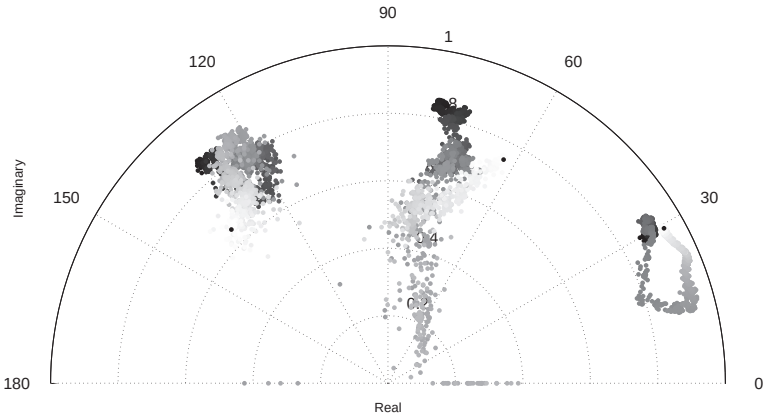
In order to process a speech signal using the AR model eqn. (3.16), the parameters, $\{a_q\}_{q \in \mathcal{Q}}$ remain to be specified. Recalling Fig. 3.7 and the discussion in sect. §3.3.2, the AR parameters specify the resonant peaks in the vocal tract response. The vocal tract is continually changing, such that the AR model in eqn. (3.16) should compensate for the time-variance of the vocal tract.

3.3.4 Time-variance of speech

To illustrate the time-varying nature of speech, consider taking a sliding window of step size $n = 1$samples, and block length $N = 250$ samples (i.e., 300msec length at sampling frequency $f_s = 8$kHz) over a 0.2s long segment of speech data from a female speaker as shown in Fig. 3.9 and described in [1]: in each window, $i = 1, \ldots, L$, $Q = 6$ stationary AR coefficients are extracted by solving the standard Yule-Walker equations [127]. Fig. 3.10a demonstrates that the overall envelope of the correspond-



(a) *Graph showing smooth variation of the parameter mean over a sliding window of* $\mathbf{a}_{2,0:t}$, $\mathbf{a}_{4,0:t}$ *and* $\mathbf{a}_{6,0:t}$.



(b) *Graph showing rapid parameter variation over large areas within the unit cirlce.*

Figure 3.10: *Pole trajectories and variation of parameters extracted from a sliding window of* 250 *samples* (0.03*s*) *block length over a real speech segment for model order* 6 *at sampling frequency* 8 *kHz as illustrated in Fig. 3.9.*

ing extracted parameters varies relatively smoothly with time. The parameters exhibit smooth variations over a short duration of around 40 samples, i.e., 0.5msec or 200Hz, after which the fluctuations in the form of spikes are visible. As the fundamental frequency of adult women is approximately $f_0 = 200$Hz, the spikes in Fig. 3.10a correspond to the glottal openings and closings in the speech signal.

The poles of the speech signal correspond to the roots of the speech parameters. Their trajectories with time are shown in Fig. 3.10b, where early samples ($t \approx 0$) are shown as light grey dots whereas late samples ($t \approx 0.2$s) correspond to black dots. The pole positions vary rapidly with relatively smooth trajectories over large areas within the unit circle. The smooth variation of pole movements is also discussed in [128]

amongst others.

As both the poles and parameters of the speech signal vary rapidly with time, modelling speech as a stationary process leads to poor approximations of the signal. Instead, speech should be modelled as a non-stationary, or time-varying, process.

Local as well as global time-variation of the signal can be captured using time-varying AR (TVAR) processes, i.e., [129]

$$x_t = \sum_{q \in \mathcal{Q}} a_{q,t}\, x_{t-q} + \sigma_{v_t}\, v_t, \tag{3.18}$$

where $\{a_{q,t}\}_{q \in \mathcal{Q}}$ is the set of Q TVAR coefficients. Based on the source *signal* model in eqn. (3.18), source *parameter* models need to be specified that capture the characteristics of source generation. As the excitation of the model has already been discussed in sect. §3.3.2, the source parameters are the only remaining unknowns in eqn. (3.18) and hence the review of parameter models concludes the scope of this chapter.

3.4 Source parameter models

Recalling Fig. 3.7 the source excitation switches from a glottal pulse waveform to turbulent noise with the closing of the vocal cords and hence the change from voiced to unvoiced sounds. Similarly, in order to structure unvoiced phonemes appropriately, the vocal tract response changes with the closing of the vocal cords. To accurately represent the vocal tract, the AR model of the speech production mechanism should account for the switching between voiced and unvoiced phonemes not only in the excitation signal but also in the vocal tract response itself. As the resonant peaks in the vocal tract response are defined by the AR parameters, $\{a_q\}_{q \in \mathcal{Q}}$, appropriate AR parameter models should be chosen for both voiced and unvoiced speech segments. Whilst voiced parameter models should enforce harmonicity in the resulting signal, unvoiced parameter models should impose turbulent noise properties corresponding to the discussion in sect. §3.3.2 on page 41 ff. Sect. §3.4.1 to sect. §3.2.2 therefore introduce parameter models for voiced and unvoiced speech relevant to the developments in this thesis. Sect. §3.4.1 discusses a stochastic TVAR model based on first-order Markov chains, applicable to unvoiced speech. Sect. §3.2.2 introduces a PFS representation of the parameters suitable for modelling of the resonant frequencies (or formants) in voiced speech.

3.4.1 Dynamic TVAR parameter model for slowly speech variation

Recalling the trajectory of speech parameters in Fig. 3.10a, the TVAR parameters vary slowly and relatively smoothly in comparison to the speech signal they are extracted from (Fig. 3.9). The smooth and slowly varying behaviour can be represented by a first-order Markov chain with low variance on the parameters [130–133]. A first-order Markov chain is a random process, where the states at time t depend directly only on the states at $t-1$. Therefore, the source parameters are expressed as:

$$\mathbf{a}_t = \mathbf{a}_{t-1} + \mathbf{\Sigma}_{\mathbf{a}_t}\,\mathbf{r}_{\mathbf{a}_t} \qquad\qquad \mathbf{r}_{\mathbf{a}_t} \sim \mathcal{N}\left(\mathbf{0}_{Q\times 1},\, \mathbf{I}_Q\right) \tag{3.19}$$

where $Q_t = Q$, $\mathbf{a}_t = \begin{bmatrix} a_{1,t} & \dots & a_{Q,t} \end{bmatrix}^T$ is the set of Q source parameters, and $\mathbf{\Sigma}_{\mathbf{a}_t} = \operatorname{diag}\begin{bmatrix} \sigma^2_{a_{1,t}} & \dots & \sigma^2_{a_{Q,t}} \end{bmatrix}$ is the covariance on the random walk.

At each time step, t, each source parameter, $a_{q,t}$, $q \in \mathcal{Q}$, is dependent only on its value at the previous time step, $a_{q,t-1}$ as well as the driving WGN of the process and its known variance, $\sigma^2_{a_{q,t}}$. This property of first-order Markov chains makes the parameter model in eqn. (3.19) particularly apt for sequential analysis and has thus been applied in several sequential Monte Carlo (SMC) frameworks, such as the work by Doucet *et al.* [24, 131, 132]. Due to the resemblance of its parameter trajectory to the slowly varying speech parameters and the applicability to SMC approaches, the dynamic TVAR parameter model is adapted in this thesis in Chaps. 6 and 7 as the basic speech model used to develop the proposed methodology for blind speech dereverberation.

An issue frequently encountered with the TVAR parameter model and circumvented in several attempts is the necessity to constrain the parameters in eqn. (3.19) to take on stable values only to establish stability of the overall speech model.

3.4.1.1 Enforcing stability of the TVAR parameters

To ensure stability for linear systems, every bounded input has to produce a generate output as established by James *et al.*in 1946 [134]. Hence, for linear time-invariant systems, the poles – i.e., the roots of the Q polynomials in eqn. (3.19) – have to lie within the unit circle [135]. If all eigenvalues of the system possess negative real parts, the system is satisfies exponential asymptotic stability in the sense of Lyapunov [136] and and is therefore bounded-input bounded-output (BIBO) stable [137]. For the linear time-varying case, the system can be *temporarily* unstable, yet lead to an *globally* stable system. Hence, stability constraints are not as obvious and can lead to ambiguities between Lyapunov and BIBO stability [138]. As shown by Anderson and Moore [137],

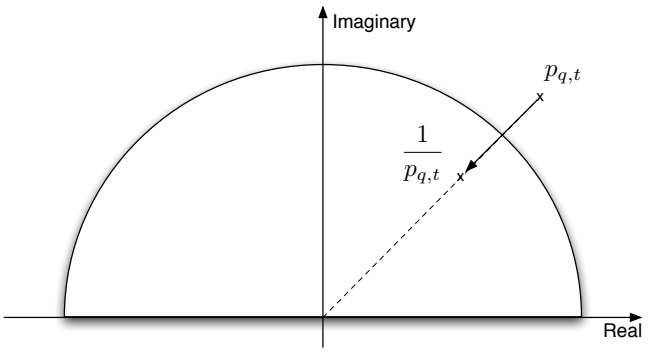


Figure 3.11: *Reflection of an unstable pole,* $p_{q,t}$*, into the unit circle,* $1/p_{q,t}$.

a bounded energy content of the input over a time-interval (rather than time-instant) is required to ensure stability of linear time-varying (LTV) systems. Hence, internal Lyapunov and external stability are equivalent [137]. However, it can be difficult to identify appropriate Lyapunov functions to analyse the asymptotic stability of the system. Hence, Juntunen *etal* [139–141] assume systems to be stable if the poles at each time instance, t, lie within the unit circle.

Stability can therefore be enforced by rejecting AR parameter sets, $\mathbf{a}_t$ whose poles, $\mathbf{p}_t$, lie on or outside the unit circle, i.e., where $|\mathbf{p}_t| \geq 1$ and where $|\cdot|$ denotes the absolute value.

The rejection rule of unstable samples is equivalent to introducing an indicator function, $\mathbb{I}_{\mathcal{A}_Q}(\mathbf{a}_t)$ over the region of support, $\mathcal{A}_Q$, of the source parameters, i.e.,

$$\mathbb{I}_{\mathcal{A}_Q}(\mathbf{a}_t) = \begin{cases} 1 & \text{if } \mathbf{a}_t \in \mathcal{A}_Q \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

The source parameters in eqn. (3.19) thus have a pdf of the form

$$p(\mathbf{a}_t \mid \mathbf{a}_{t-1}) \propto \mathcal{N}\left(\mathbf{a}_t \,\middle|\, \mathbf{a}_{t-1}, \boldsymbol{\Sigma}_{\mathbf{a}_t}\right) \mathbb{I}_{\mathcal{A}_Q}(\mathbf{a}_t) \tag{3.21}$$

where the initial state of the chain is $p(\mathbf{a}_0) \propto \mathcal{N}\left(\mathbf{a}_0 \,\middle|\, \mathbf{0}_{Q\times 1}, \boldsymbol{\Sigma}_{\mathbf{a}_0}\right) \mathbb{I}_{\mathcal{A}_Q}(\mathbf{a}_0)$, and where the Markov parameters $\boldsymbol{\Sigma}_{\mathbf{a}_t}$ for $t \geq 0$ are assumed known.

To avoid rejection of samples through eqn. (3.21), stability can be enforced by reflecting unstable poles back into the unit circle [142] as illustrated Fig. 3.11, i.e.,

$$p_{q,t} = \frac{1}{p_{q,t}}. \tag{3.22}$$

Data: Vector of poles, $\hat{\mathbf{p}} = \begin{bmatrix} \hat{p}_1 & \dots & \hat{p}_Q \end{bmatrix}^T$
Result: Vector of stable poles, $\mathbf{p} = \begin{bmatrix} p_1 & \dots & p_Q \end{bmatrix}^T$
for $q = 1, \dots, Q$ **do**
  if $|\hat{p}_q| \geq 1$ **then**
1 Reflect the pole back into the unit circle: $p_q = 1/\hat{p}_q$ (eqn. (3.22));
  else
2 Poles remain unchanged: $p_q = \hat{p}_q$;
  end
end

Algorithm 3.1: $\mathbf{p}$ = CheckStability($\hat{\mathbf{p}}$)

where $\mathbf{p}_t \triangleq \begin{bmatrix} p_{1,t} & \dots & p_{Q,t} \end{bmatrix}^T$ are the Q poles corresponding to the roots of $\mathbf{a}_t$. As poles appear in complex-conjugate pairs, the reflection changes the radius but leaves the phase unchanged.

The reflection of poles into a stable region changes the probability distribution of the TVAR parameters. Consider the sample drawn from eqn. (3.19) as an auxiliary sample subject to verification of its stability:

$$\hat{\mathbf{a}}_t = \mathbf{a}_{t-1} + \boldsymbol{\Sigma}_{\mathbf{a}_t} \mathbf{r}_{\mathbf{a}_t}, \qquad \mathbf{r}_{\mathbf{a}_t} \sim \mathcal{N}(\mathbf{0}_{Q\times 1}, \mathbf{I}_Q).$$

or, in other words:

$$p(\hat{\mathbf{a}}_t \mid \mathbf{a}_{t-1}) = \mathcal{N}\left(\hat{\mathbf{a}}_t \,\middle|\, \mathbf{a}_{t-1}, \boldsymbol{\Sigma}_{\mathbf{a}_t}\right) \tag{3.23}$$

The stability of the source parameter is ensured by evaluating

$$\mathbf{a}_t = \begin{cases} \hat{\mathbf{a}}_t & \text{if } \hat{\mathbf{a}}_t \in \mathcal{A}_Q \\ f(\hat{\mathbf{a}}_t) & \text{otherwise} \end{cases} \tag{3.24}$$

where $f(\cdot)$ corresponds to the translation in eqn. (3.22) and is a one to one mapping where real poles are mapped to real poles, $f : \mathbb{R} \to \mathbb{R}$, whilst complex poles remain complex, $f : \mathbb{C} \to \mathbb{C}$.

The reflection of unstable parameters is summarised in Alg. 3.1 and the dynamic TVAR parameter model using the reflection of unstable parameters is summarised in Alg. 3.2. Due to the reduction in computational waste as compared to the introduction of an indicator function in eqn. (3.20), the reflection of TVAR parameters is used in Chaps. 6 and 7 to ensure stable TVAR parameters.

Initialisation:
for $t = 0, \ldots, Q-1$ **do**
1 Initialise the source signal, e.g., $x_t = 0$;
2 Initialise the source parameters, i.e., $\hat{\mathbf{a}}_t \sim \mathcal{N}(\mathbf{0}_{Q\times 1}, \boldsymbol{\Sigma}_{\mathbf{a}_0})$ and compute the poles, $\hat{\mathbf{p}}_t$;
3 Ensure stability of poles: $\mathbf{p}_t = \text{CheckStability}(\hat{\mathbf{p}}_t)$ (Alg. 3.1) and compute the parameters, $\mathbf{a}_t$;
end
for $t \geq Q$ **do**
4 Draw the auxiliary source parameter sample $\hat{\mathbf{a}}_t \sim \mathcal{N}(\mathbf{a}_{t-1}, \boldsymbol{\Sigma}_{\mathbf{a}_t})$ (eqn. (3.23)) and compute the poles $\mathbf{p}_t$;
5 Ensure stability of poles: $\mathbf{p}_t = \text{CheckStability}(\hat{\mathbf{p}}_t)$ (Alg. 3.1) and compute the parameters, $\mathbf{a}_t$;
6 Generate the synthetic speech signal:

$$x_t = \sum_{q\in\mathcal{Q}} a_{q,t}\, x_{t-q} + \sigma_{v_t}\, v_t. \tag{3.18}$$

end

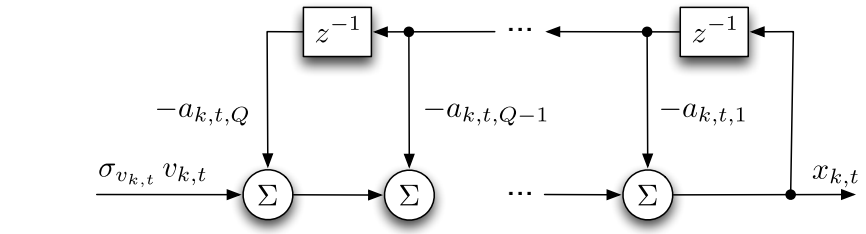
Algorithm 3.2: Dynamic TVAR parameter model

Another alternative for ensuring stability of the model was proposed by Fong and Godsill [143], where the TVAR source model is reparameterised in terms of PARCOR coefficients rather than TVAR parameters. This approach is also utilised in Chap. 8 for an improved speech model.

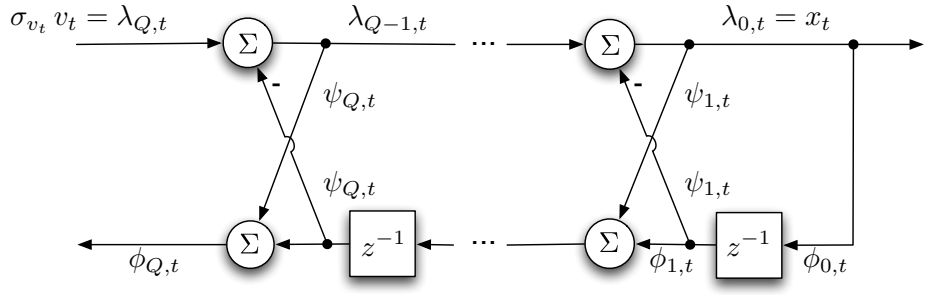
3.4.2 PARCOR representation of the AR parameters for ensured stability

The TVAR model in eqn. (3.18) on page 48 is usually represented using a direct-form IIR filter structure as illustrated in Fig. 3.12a. As discussed in, e.g., [144, ch. 9.3.5], an equivalent representation is the lattice structure as illustrated in Fig. 3.12b, generating the signal by means of a forward and a backward lattice that are connected by a sequence of reflection coefficients. Thus, lattice structures are parameterised by a sequence of reflection coefficients rather than AR parameters. The reflection coefficients are also known as PARCOR coefficients as discussed in sect. §3.4.2.1.

The incentive to choose lattice structures over the direct-form representation is two-fold: Firstly, PARCOR coefficients by definition lie between -1 and 1. Due to these boundary conditions and their relation to the AR parameters to be discussed below, any valid choice of PARCOR coefficients corresponds to *stable* AR parameters (see sect. §3.4.2.1). Secondly, the reflection coefficients are identical in form to the reflection



(a) *Direct-form structure*



(b) *Lattice structure*

Figure 3.12: *Equivalent lattice and direct-form IIR structures [144]*

coefficients of the acoustic tube model describing the propagation and reflection of a sound wave in the concatenated tubes representing the different sections of the vocal tract (see sect. §3.4.2.2). Lattice structures are therefore extensively used in speech processing and adaptive filtering applications [107, 142, 144, 145].

The speech signal generated by the lattice structure in Fig. 3.12b is expressed in terms of the last forward-stage output, i.e.,

$$x_t = \lambda_{0,t} = \phi_{0,t}, \tag{3.25}$$

where the lattice stage outputs, $\lambda_{0,t}$ to $\lambda_{Q,t}$ are defined recursively via

$$\lambda_{Q,t} = \phi_{Q,t} = \sigma_{v_t}\, v_t \tag{3.26a}$$

$$\lambda_{q-1,t} = \lambda_{q,t} - \psi_{q,t}\, \phi_{q-1,t-1} \qquad \text{for } q = 1, \ldots, Q \tag{3.26b}$$

$$\phi_{q,t} = \psi_{q,t}\, \lambda_{q-1,t} + \phi_{q-1,t-1} \qquad \text{for } q = 1, \ldots, Q. \tag{3.26c}$$

where v_t is the excitation of the filter with variance $\sigma^2_{v_t}$, i.e., the turbulent noise from the lungs. In this dissertation, reflection coefficients are utilised in Chap. 8 for the representation of resonator circuits for a formant synthesis speech model (see sect. §3.4.3). As resonator circuits are second-order filters, the following discussion is therefore limited to $Q = 2$. The relation between AR and reflection coefficients are thoroughly discussed for the general case in, e.g., [142, 144].

Solving the recursions eqn. (3.26) for for the second-order case, i.e., $Q = 2$, and inserting into eqn. (3.25), the lattice signal model reduces to (see Appendix A.4.1)

$$x_t = -\psi_{1,t}\,(1 + \psi_{2,t})\,x_{t-1} - \psi_{2,t}\,x_{t-2} + \sigma_{v_t}\,v_t \tag{3.27}$$

Comparing to a second-order TVAR speech model in eqn. (3.18) on page 48, i.e.,

$$x_t = -a_{1,t}\,x_{t-1} - a_{2,t}\,x_{t-2} + \sigma_{v_t}\,v_t, \tag{3.18}$$

the lattice and direct-form structures differ only the definition of their parameters. By comparing eqns. (3.27) and (3.18) the relation between the TVAR and reflection coefficients can be expressed as:

$$a_{1,t} = \psi_{1,t}\,(1 + \psi_{2,t}) \tag{3.28a}$$

$$a_{2,t} = \psi_{2,t}. \tag{3.28b}$$

Due to the direct relation between $\mathbf{a}_t$ and $\boldsymbol{\psi}_t$ in eqn. (3.28), any constraints on the reflection coefficients directly translate to the TVAR parameters. As mentioned above, the reflection coefficients can also be interpreted as PARCOR coefficients and therefore, by definition are bounded between ± 1. Sect. §3.4.2.1 derives the partial correlation interpretation of the reflection coefficients and shows how the resulting constraint enforces stability on the TVAR parameters. The relation between the lattice structure reflection coefficients and the reflection coefficients are discussed in sect. §3.4.2.2.

3.4.2.1 Bounds of the reflection coefficients

The lattice recursions in eqns. (3.26b) and (3.26c) can be written by slight rearrangement of eqn. (3.26b) in the form of an autoregression, i.e.,

$$\begin{aligned}\lambda_{q,t} &= \lambda_{q-1,t} + \psi_{q,t}\,\phi_{q-1,t-1} = \lambda_{q-2,t} + \psi_{q-1,t}\,\phi_{q-2,t-1} + \psi_{q,t}\,\phi_{q-1,t-1} = \dots \\ &= \phi_{0,t} + \psi_{1,t}\,\phi_{0,t-1} + \cdots + \psi_{q-1,t}\,\phi_{q-2,t-1} + \psi_{q,t}\,\phi_{q-1,t-1}.\end{aligned} \tag{3.29}$$

As an expression for neither $\lambda_{q,t}$ or $\phi_{q,t}$ is available, an expression $\psi_{q,t}$ cannot be found by simply rearranging eqn. (3.29). Instead, an expression for the reflection coefficients can be found by maximising the variance between the forward and backward stage output, $\lambda_{q,t}$ and $\phi_{q,t}$, and solving for the reflection coefficients as demonstrated in Appendix A.4.2. The lattice reflection coefficients can hence be expressed as

$$\psi_{q,t} = \frac{\operatorname{cov}\left[\lambda_{q-1,t},\ \phi_{q-1,t-1}\right]}{\operatorname{var}\left[\phi_{q-1,t-1}\right]} \tag{3.30}$$

Eqn. (3.30) is also known as the partial correlation (PARCOR) function between $\lambda_{q-1,t}$ and $\phi_{q-1,t}$. As the term PARCOR coefficient is generally preferred in the literature, the following will refer to the reflection coefficients as the PARCOR coefficients.

By definition, the correlation between two random variables is bound between ± 1 [146]. Therefore, the reflection coefficients obey at all times

$$-1 \leq \psi_{q,t} \leq 1. \tag{3.31}$$

Inserting into eqn. (3.28), the source parameters thus take extrema between

$$a_{1,t} = \begin{cases} 2, & \text{if } \psi_{1,t} = 1 \text{ and } \psi_{2,t} = 1 \\ -2, & \text{if } \psi_{1,t} = -1 \text{ and } \psi_{2,t} = 1 \end{cases} \tag{3.32a}$$

$$a_{2,t} = \begin{cases} 1, & \text{if } \psi_{2,t} = 1 \\ -1, & \text{if } \psi_{2,t} = -1 \end{cases} \tag{3.32b}$$

As discussed in [146, Chapt. 3.2.4], stability is enforced if $-2 \leq a_{1,t} \leq 2$ and $-1 \leq a_{2,t} \leq 1$, such that the boundary conditions of the TVAR parameters in eqns. (3.32a) and (3.32b) are stable. Therefore, the PARCOR coefficients guarantee stable parameters due to the non-linear relationship between $\boldsymbol{\psi}_t$ and $\mathbf{a}_t$. The stability of TVAR parameters will be elaborated on sect. §3.4.3.4 on page 61.

3.4.2.2 Relation of the PARCOR model to the acoustic tube model

Recalling the discussion in sect. §3.3.1 on page 38, the acoustic tube model represents the vocal tract as a concatenation of lossless tubes of equal tubes. At the junctions between tube segments, part of sound waves is propagated, whilst part of it is reflected. Propagated and reflected waves are related by the reflection coefficient of the acoustic tube as illustrated in Fig. 3.5 on page 39. This figure triggers particular interest in the discussion of PARCOR coefficients due to its resemblance to a lattice structure. As both the PARCOR model as well as the vocal tract model are represented by a lattice structure, the question arises to what extent the two representations are equivalent. The similarities between both models can be illuminated by investigation of their transfer functions.

The transfer function, $V(z)$, of the PARCOR model can be obtained by taking the z-transform of the difference equation in eqn. (3.27), such that

$$V^{\text{PARCOR}}(z) = \frac{N_P(z)}{A(z)} \tag{3.33}$$

where $A(z) = 1 + (\psi_{1,t}\,\psi_{2,t} + \psi_{1,t-1})z^{-1} + \psi_{2,t}\,z^{-2}$ and $N_P(z)$ is the z-transform of the variance of the PARCOR process. By slightly modifying the forward- and backward-stages in eqn. (3.26) in the z-domain, the denominator of the transfer function in eqn. (3.33) can be expressed in terms of the following recursions (see Appendix A.4.3):

$$A(z) = A_Q(z) \tag{3.34a}$$

$$A_0(z) = 1 \tag{3.34b}$$

$$A_q(z) = A_{q-1}(z) + \psi_{q,t}\,z^{-q}\,A_{q-1}(z^{-1}), \qquad q = 1, \ldots, Q \tag{3.34c}$$

As the transfer function of the PARCOR structure in Fig. 3.12b can be expressed in terms of a recursion on the denominator, the aim is to investigate whether the transfer function of the acoustic tube model can be written as a recursion as well.

Comparing the recursions describing the transfer function of the vocal tract in eqn. (3.8) on page 40 to those describing the transfer function of the PARCOR model in eqn. (3.34), the $D_k(z)$ and $A_q(z)$ are equivalent and hence the transfer function of the vocal tract in eqn. (3.10) on page 41 is equivalent to that of the lattice structure in eqn. (3.33). Therefore, representing the TVAR model in terms of its reflection (or PARCOR) coefficients directly relates the model to the propagation and reflection of sound within the vocal tract.

As the recursion $D_k(z)$ is equivalent to $A_q(z)$ and, by comparison of eqns. (3.8c) and (3.34c), the reflection coefficients of the PARCOR model, $\boldsymbol{\psi}_t$, are equivalent to the reflection coefficients of the acoustic tube model, r_k, the PARCOR coefficients can hence be expressed of the form

$$\psi_{q,t} = r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}, \tag{3.35}$$

where A_k is the area of the acoustic tube segment, $k \in \mathcal{K}$.

Although the dynamic TVAR parameter model presented in this section simulate the slowly varying nature of speech parameters and are particularly suitable for sequential processing, other physical properties of voiced speech, such as harmonic or even sinusoidal components, are not reflected in the parameters. Therefore, similar to the reparameterisation of the TVAR model in terms of PARCOR coefficients in [143], Beierholm and Winther [111] propose to reparameterise the TVAR model in terms of resonant frequencies, bandwidths, and gains. This parameterisation leads to an interpretation of formant synthesisers (see sect. §3.2.2) from a TVAR perspective.

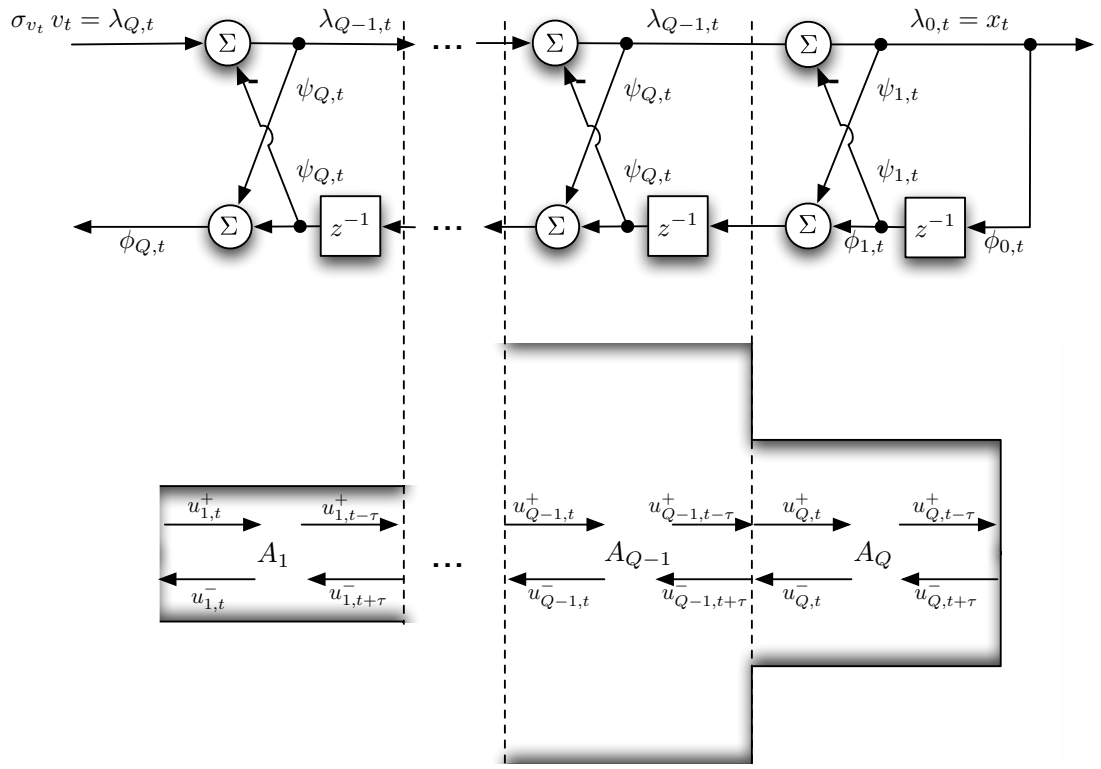


Figure 3.13: *Modelling of the vocal tract by means of IIR lattice structures. Each tube segment is equivalent to a stage in the lattice structure, where the reflection coefficients describe the ratio between the propagated and reflected sound wave.*

3.4.3 Parallel formant synthesis from a TVAR perspective

As explained in sect. §3.2.2, parallel formant synthesisers model the formants (or spectral peaks) generated in the vocal tract as a parallel concatenation of resonators, each of which models one formant. A resonator, in physical terms, can be any body or device to which a vibrating column such as a string, air column, or membrane is attached. The resonator acts as a filter on the harmonics of the vibrating column such that some harmonics are amplified and some are attenuated. Resonators therefore oscillate at certain frequencies, i.e., the resonant frequencies, with higher amplitudes than at other frequencies.

From a digital perspective, resonators are implemented by means of second order TVAR processes with a complex-conjugate pair located inside and close to the unit circle. Hence, the TVAR speech signal model in eqn. (3.18) on page 48 is slightly modified. As K resonators are connected in parallel in the PFS to model K formants, K TVAR models of order $Q = 2$ are employed, where for each resonator $k \in \mathcal{K}$, the

corresponding TVAR model is given by

$$x_{t,k} = a_{1,t,k}\, x_{t-1,k} + a_{2,t,k}\, x_{t-2,k} + \sqrt{g_{t,k}}\, v_t, \qquad v_t \sim \mathcal{N}(0,\, 1)\,. \tag{3.36}$$

where $x_{t,k}$ is the output and $a_{1,t,k}$ and $a_{2,t,k}$ the source parameters of the k^{th} resonator, and $g_{t,k}$ is the resonator gain, equivalent to the excitation noise, $\sigma^2_{v_t}$ in eqn. (3.18). The K resonator signals are combined by summing over all filter outputs, i.e.,

$$x_t = \sum_{k \in \mathcal{K}} x_{t,k}. \tag{3.37}$$

Effectively, the source signal model is therefore only mildly modified in that the output signal of K TVAR models is combined. The source parameters of the PFS model are dictated by the resonant properties of the resonator circuit, i.e., by the resonator frequency, bandwidth and gain. Sect. §3.4.3.1 to sect. §3.4.3.4 therefore discuss the TVAR parameter model specified in terms of the resonant frequency, bandwidth and gain. Sect. §3.4.3.1 demonstrates that the design specifications of the resonator frequency response is expressed in terms of the resonant frequency, bandwidth, and gain and can be directly related to the poles of the TVAR model in eqn. (3.36). Sect. §3.4.3.2 to sect. §3.4.3.3 relate the poles of the model to the resonant frequency and bandwidth. Sect. §3.4.3.4 relates the TVAR parameters to their poles, closing the loop between the relation of resonant parameters, TVAR poles, and TVAR parameters.

3.4.3.1 Relation of the parameters and resonant frequency

The TVAR parameters of each resonator, $\mathbf{a}_{t,k} = \begin{bmatrix} a_{1,t,k} & a_{2,t,k} \end{bmatrix}^T$, are calculated so as to satisfy the design criteria of the corresponding resonators. The main concern for designing the resonators is to ensure poles located near the unit circle to generate large magnitude responses at the corresponding positions in the spectrum. Five design equations are generally used to specify the properties of the filter [147]:

$$|H_{t,k}(0)| = G_0 \tag{3.38a}$$

$$|H_{t,k}(\pi)| = G_\pi \tag{3.38b}$$

$$\frac{\partial}{\partial \omega} |H_{t,k}(\omega)| \Big|_{\omega = \omega_{t,k}} = 0 \tag{3.38c}$$

$$|H_{t,k}(\omega_{t,k})| = G_R \tag{3.38d}$$

$$|H_{t,k}(\omega) \Big|_{\omega = \omega_{t,k} \pm B/2}| = G_B, \tag{3.38e}$$

where $H_{t,k}(\omega)$ is the frequency response of the filter, G_0 is the gain at direct current (DC), G_π is the gain at Nyquist frequency, G_R is the gain at resonance, G_B is the gain at

the 3dB bandwidth, B is the 3dB bandwidth, $\omega = {}^{2\pi f}/_{f_s}$ denotes the radial frequency, $\omega_{t,k}$ is the radial frequency at resonance, and f_s is the sampling frequency. G_0 and G_π are typically chosen to be equal and are set to 0dB, i.e., $G_0 = G_\pi = 1$ in order to facilitate the cascading of several parametric equaliser filters [147]. The frequency response, $H_{t,k}(\omega)$, of any second-order AR process can be expressed in terms of its poles by

$$\begin{aligned} H_{t,k}(\omega) &= \frac{g_{t,k}}{(1 - p_{1,t,k}\, e^{-\jmath\,\omega})\,(1 - p_{2,t,k}\, e^{-\jmath\,\omega})} \\ &= \frac{g_{t,k}}{(1 - r_{t,k}\, e^{\jmath\phi_{t,k}}\, e^{-\jmath\,\omega})\,(1 - r_{t,k}\, e^{-\jmath\phi_{t,k}}\, e^{-\jmath\,\omega})} \end{aligned} \tag{3.39}$$

where $\mathbf{p}_{t,k} \triangleq \begin{bmatrix} p_{1,t,k} & p_{2,t,k} \end{bmatrix}^T$ is the set of poles, $r_{t,k}$ is the pole radius, and $\phi_{t,k}$ is the pole phase, and where $p_{1,t,k} = r_{t,k}\, e^{\jmath\phi_{t,k}}$. Note that complex poles occur in complex conjugate pairs, such that $p_{2,t,k} = p_{1,t,k}^\star$, where $(\cdot)^\star$ denotes the complex conjugate. The poles can therefore be related to the resonant frequency, bandwidth, and gain by inserting the design equations in eqn. (3.38) into eqn. (3.39).

3.4.3.2 Relating the poles to the resonant frequency

As a resonance is a local peak in the magnitude response of the filter, the resonant frequency is the frequency that maximises the magnitude spectrum of eqn. (3.39) as described by the design specification in eqn. (3.38c). Thus, by taking the derivative of the magnitude of $H_{t,k}(\omega)$ at the resonant frequency $\omega = \omega_{t,k}$, and solving for $f_{t,k} = {}^{\omega_{t,k}\, f_s}/_{2\pi}$, the resonant frequency can be expressed as (see Appendix A.5.2)

$$f_{t,k} = \frac{f_s}{2\pi} \arccos\left(\frac{1 + r_{t,k}^2}{2 r_{t,k}} \cos\phi_{t,k} \right). \tag{3.40}$$

$f_{t,k}$ can therefore be fully expressed in terms of the TVAR pole radius and phase. In order to simplify the expression in eqn. (3.40), it is often assumed that the pole radius is very close to unity, i.e., $r_{t,k} \approx 1$ (see, e.g., [144]). Therefore, $\phi_{t,k} \approx {}^{2\pi f_{t,k}}/_{f_s}$.

3.4.3.3 Relating the poles to the resonant bandwidth

The bandwidth, $B_{t,k}$, of low-order filters is determined by the difference between the band-edge frequencies at an attenuated level of 3dB relative to the maximum of the frequency response [148] (see Fig. 3.14). The bandwidth of each resonator is thus

$$B_{t,k} = \omega_u - \omega_\ell, \tag{3.41}$$

where ω_u and ω_ℓ are the upper and lower band-edge frequencies respectively. The band-edge frequencies are measured at the 3dB point of the magnitude response, i.e.,

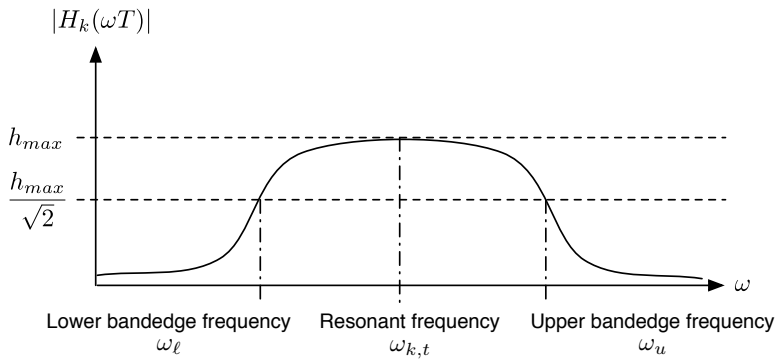


Figure 3.14: *Magnitude response showing upper and lower band-edge frequencies, ω_u and ω_ℓ; 3dB bandwidth is the difference between the band-edge frequencies.*

at $h_{max}/\sqrt{2}$ where h_{max} is the maximum of the magnitude response. Thus, consulting Fig. 3.14, the magnitude response at the band-edge frequencies must equal the magnitude response at an attenuated level -3dB from the frequency response at resonance:

$$|H_{t,k}(\omega_u)| = |H_{t,k}(\omega_\ell)| = \frac{|H_{t,k}(\omega_{t,k})|}{\sqrt{2}} \tag{3.42}$$

The band-edge frequencies, and hence the bandwidth via eqn. (3.41) can therefore be identified by inserting $H_{t,k}(\omega)$ into eqn. (3.42) and solving for ω_u and ω_ℓ. Hence, following the derivations in Appendix A.5.3:

$$\omega_{\{u,\ell\}} = \arccos\left\{\frac{1}{2r_{t,k}}\left((1+r_{t,k}^2)\cos\phi_{t,k} \pm (1-r_{t,k}^2)\sin\phi_{t,k}\right)\right\}. \tag{3.43}$$

The band-edge frequencies, and, by inserting eqn. (3.43) into eqn. (3.41), the bandwidth of the resonator can thus be expressed completely by the pole radius and phase.

Recall that the overall aim of this section is to derive a mapping from the known resonant frequency, bandwidth, and gain of each resonator to the unknown TVAR parameters of the speech model to establish a TVAR model of the parallel formant synthesiser. As the model order of each resonator is $Q = 2$, only two TVAR parameters need to be determined, corresponding to only two pole positions. As the pole positions are complex conjugate pairs, only two unknowns need to be identified, namely the pole radius and phase. Given the pole radius and phase, the TVAR parameters can be thus be fully identified. As the resonant frequency in eqn. (3.40) and the bandwidth in eqns. (3.41) and (3.43) are fully specified in terms of the radius and phase. Hence, the pole positions and thus the TVAR parameters can be solved using eqns. (3.40), (3.41) and (3.43). The resonator gain is thus not required for the solution of the TVAR parameters. In fact, recalling eqns. (3.36) and (3.39), the gain is equivalent to the standard deviation of the signal model. Thus, assuming a pre-specified resonator

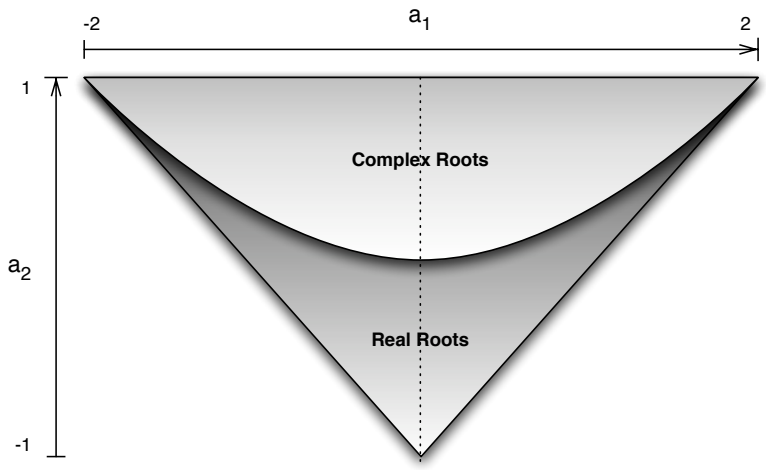


Figure 3.15: *Admissible region of stable second-order AR parameters after [146].*

gain, the process excitation variance of the speech model is given as well.

Hence, having specified the poles in terms of the resonant frequency and bandwidth and the process excitation variance in terms of the resonant gain, it remains to relate the poles to the TVAR parameters in order to fully specify the TVAR PFS speech model in eqn. (3.36) on page 58.

3.4.3.4 Relating the AR parameters to their poles

The poles, $\mathbf{p}_{t,k}$ of the system can therefore be related via the frequency response in eqn. (3.39) to the resonant frequency, bandwidth, and gain, $f_{t,k}$, $B_{t,k}$ and $g_{t,k}$ through the design specifications in eqn. (3.38). Taking the z-transform of the impulse response in eqn. (3.36), $H_{t,k}(\omega)$ can also be expressed in terms of the TVAR coefficients, i.e.,

$$H_{t,k}(\omega) = \frac{g_{t,k}}{1 + a_{1,t,k}\, z^{-1} + a_{2,t,k}\, z^{-2}}. \tag{3.44}$$

Combining eqns. (3.44) and (3.39), the well-known relationship can be derived between TVAR coefficients and the pole radius and phase of the resonator, defining the peak the in the spectral response. The necessary derivations can be found in, e.g., [144] and are summarised for completeness in Appendix A.5.1, such that, for complex poles:

$$a_{1,t,k} = -2r_{t,k}\cos\phi_{t,k} \qquad\qquad a_{2,t,k} = r_{t,k}^2 \tag{3.45}$$

such that $0 \leq a_{2,t,k} \leq 1$ and $-2 \leq a_{1,t,k} \leq 2$ in order enforce stable poles, where $0 \leq r_{t,k} < 1$. Note that this relation reduces to $a_{1,t,k} = -(r_{t,k}+r^\star_{t,k})$ and $a_{2,t,k} = r_{t,k}\, r^\star_{t,k}$ for real poles, where $r^\star_{t,k}$ denotes the radius of the second pole. For real poles, the

1 Initialise the source signal, e.g., $\{x_{t,k} = 0 : k \in \mathcal{K},\ t = 0, 1\}$;

for *each sample* $t \geq 2$ **do**

for *each resonator* $k = 1, \ldots, K$ **do**

2 Specify the fundamental frequency, $f_{t,k}$, bandwidth, $B_{t,k}$, and gain, $g_{t,k}$;

3 Determine the band-edge frequencies, ω_u and ω_ℓ based on the resonant frequency and using the specified bandwidth: $B_{t,k} = \omega_u - \omega_\ell$ (3.41);

4 Solve for the radius and phase, $r_{t,k}$, and $\phi_{t,k}$ of the poles using the resonant frequency and the band-edge frequencies:

$$f_{t,k} = \frac{f_s}{2\pi} \arccos\left(\frac{1 + r_{t,k}^2}{2r_{t,k}} \cos\phi_{t,k}\right) \tag{3.40}$$

$$\omega_{\{u,\ell\}} = \arccos\left\{\frac{1}{2r_{t,k}}\left((1 + r_{t,k}^2)\cos\phi_{t,k} \pm (1 - r_{t,k}^2)\sin\phi_{t,k}\right)\right\}. \tag{3.43}$$

5 Using the pole radius and phase, compute the TVAR parameters: $a_{1,t,k} = -2r_{t,k}\cos\phi_{t,k}$ and $a_{2,t,k} = r_{t,k}^2$ (3.45);

6 Generate the resonator signal:

$$x_{t,k} = a_{1,t,k}\, x_{t-1,k} + a_{2,t,k}\, x_{t-2,k} + g_{t,k}\, v_t, \qquad v_t \sim \mathcal{N}(0,\ 1)\,. \tag{3.36}$$

end

7 Combine the resonator signals to generate the synthetic speech signal:

$$x_t = \sum_{k \in \mathcal{K}} x_{t,k}. \tag{3.37}$$

end

Algorithm 3.3: PFS based TVAR model

region of support of stable parameters is therefore given as $-2 \leq a_{1,t,k} \leq 2$ and $-1 \leq a_{2,t,k} \leq 1$. As $a_{2,t,k}$ for complex poles in eqn. (3.45) is quadratic in the pole radius, the region of convergence of $a_{2,t,k}$ is a parabola shape. Similarly, as $a_{2,t,k}$ is linear in $r_{t,k}$ for real poles, such that is admissible region is of triangular shape. The admissible regions for both real and complex poles are shown in Fig. 3.15.

To summarise, PFSs consist of K resonators connected in parallel, each of which models one formant and is preceded by an amplitude control of the spectral peak as mentioned in sect. §3.2.2. As discussed in sect. §3.2.2 and [102], three to five formants are generally visible in the spectrum of human speech, such that $K = 5$. The formant frequency, bandwidth and gain of each resonator can thus be assumed known. Each resonator signal can be modelled by the TVAR process in eqn. (3.36) on page 58. The resonator signals are combined to form the speech signal using eqn. (3.37) on page 58. In order to fully specify the TVAR model of each resonator, the TVAR parameters in

eqn. (3.36) need to be related to the known formant frequency and bandwidth, whilst the process excitation standard deviation is given by the formant gain. The formant frequency and bandwidth are related to the TVAR poles – i.e., the roots of the parameters – through the frequency response of the resonator via eqns. (3.40), (3.41) and (3.43). The TVAR poles can thus be specified from the known formant frequency and bandwidth. The resulting poles are related to the TVAR parameters via eqn. (3.45), such that the TVAR parameters are identified and the TVAR model in eqn. (3.36) is fully specified. The source signal generation using the PFS based TVAR speech model is summarised in Alg. 3.3.

The PFS based TVAR speech model incorporates physical information about the human speech production mechanism in the TVAR speech signal model obtained from the formant synthesis speech system model in sect. §3.2.2. As prior knowledge about the production of speech in the vocal tract is incorporated, the PFS based TVAR model leads to improved modelling as compared to the dynamic TVAR parameter model in sect. §3.4.1 [111]. The dynamic TVAR parameter model used for deriving the proposed speech dereverberation algorithm in Chaps. 6 and 7 is therefore replaced by the PFS based TVAR model in Chap. 8, demonstrating improved modelling, particularly of voiced and transient sounds.

3.5 Summary

In order to recover the clean speech signal from reverberant and noisy observations, it is advantageous to utilise information available about the speech production system and the distorting channel to remove the degrading effects of the environment. In this chapter, models of the speech production mechanism relevant to this thesis were presented. Speech models in the literature and relevant to this thesis were introduced.

It was discussed that speech models can generally be divided in speech *system* models, describing the production mechanism of speech in the vocal tract, and speech *signal* models, describing properties of the signal directly. In the discussion of speech system models, it was highlighted that the vocal tract can be modelled by a concatenation of lossless acoustic tubes to represent the different cavities in the vocal tract. Alternatively, formant synthesisers represent speech by modelling each of the speech formants as a resonator circuit. Although speech system models are not directly applied to speech in this thesis, their concepts are applied and incorporated to improve upon the presented signal models for more physically meaningful modelling of speech.

By investigating the transfer function of the acoustic tube model, an equivalent speech *signal* model was developed, modelling the speech signal as a TVAR process

excited by WGN to produce unvoiced speech resembling turbulent noise and glottal pulse waveforms to generate voiced speech resembling harmonic waves. The dynamic properties of the TVAR signal are specified by underlying *parameter models*.

By investigation of the TVAR parameters extracted from a real speech sequence, it was found that the parameters vary relatively smoothly and slowly with time and contain periodic components. Based on the smoothness and slow variation of the parameters, a dynamic TVAR parameter model was introduced, modelling the parameters as a first-order Markov chain. It was pointed out that it is desirable to incorporate prior physical knowledge about the speech production system to improve the modelling of speech. Revisiting the concept of PFSs and their representation as a concatenation of resonator circuits, a parameter model based on a TVAR interpretation of the PFS was introduced. It was shown that the TVAR parameters of the generic TVAR signal model can be related to the formant frequencies, bandwidths, and gains of the resonator circuits. Assuming that the formant frequencies, bandwidths and gains are known, the TVAR parameters can be fully specified, thus facilitating an implementation of the PFS in a TVAR model. The dynamic TVAR parameter model and the PFS model are implemented and their modelling performance compared in the speech dereverberation framework of this thesis in Chap. 6 to Chap. 8.

Based on the observation that the speech parameters contain periodic components and voiced speech is highly harmonic, a sinusoidal parameter model was discussed, considering the signal as a linear combination of sinusoids. A more generalised approach, modelling the parameters as a linear combination of basis function was also introduced. Although neither the sinusoidal model nor the basis function model are directly implemented in the dereverberation framework in this thesis, an future work extension to the work presented in this thesis proposes to incorporate the sinusoidal model in a Markov switching model combining the dynamic TVAR parameter model, the PFS model, and the sinusoidal model for modelling unvoiced, transient, and voiced speech components for more versatile speech modelling. The basis function model is pertinent in Chap. 9, where an extension to dereverberation of moving speakers is presented. The results of the proposed algorithm are compared to a similar approach to blind dereverberation of speech from moving speakers in the literature, utilising the basis function source model.

This chapter therefore provided the necessary background of the production of speech signals and their models required in this dissertation. In order to complete the discussion of speech and dereverberation, room acoustics and appropriate models are reviewed in the following chapter.

Chapter 4

Room transfer function and its models

4.1 Introduction

Room impulse responses (RIRs) can be divided into three parts: the direct path response, early reflections and late reflections. The RIR of a $2.78 \times 4.68 \times 3.2$ office with reverberation time $T_{60} = 0.2$s is illustrated in Fig. 4.1. Due to the small delay, early reflections are not perceived separately from the direct path signal and reinforce the direct signal, often referred to as the precedence effect [149, 150]. However, due to their sparse nature and the short delay between the early sound components, early reflections cause spectral colouration to the direct path signal. Late reflections are continuous in nature and are perceived as either separate echoes or reverberation, thus *impairing* intelligibility [7, 151]. Both a distinct echo as well as a succession of reflections cause a characteristic change of timbre of the direct path sound, also referred to as colouration [4]. The effects of reverberation are especially pertinent for non-native listeners [152] and listeners with impaired hearing [153].

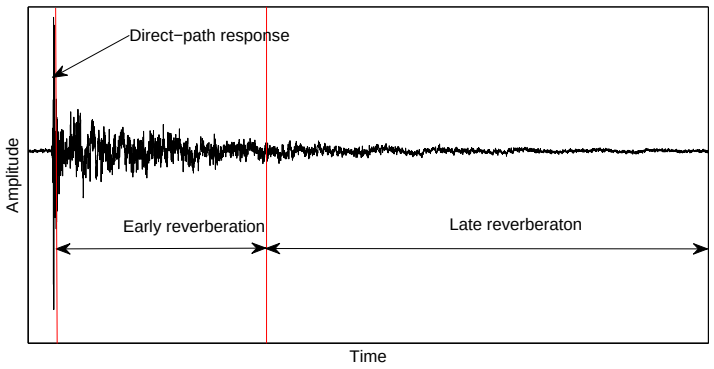


Figure 4.1: *Illustration of the RIR of a* $2.78 \times 4.68 \times 3.2$ *office with reverberation time* $T_{60} = 0.2s$.

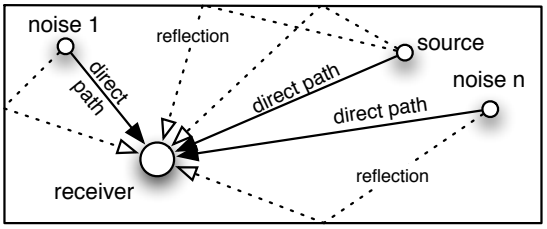


Figure 4.2: *Distant noise source filtered through separate channels*

In addition to reverberation, the signal is often distorted by interfering noise sources as well. An illustration of the distortion of speech radiated in enclosed spaces is shown in Fig. 4.2. Similar to Chap. 3, mathematical models are required to characterise and abstract the various stages involved in the distortion of speech in acoustic environments. Whilst speech models are broadly classified into speech *production* and speech *signal* models, room acoustical models can be divided into models that *simulate* the RIR and models that *describe* the RIR. Sect. §4.2 therefore discusses the general expression for room transfer functions (RTFs). A method used for simulating RIRs and known as the image-source method (ISM) is discussed in sect. §4.3. The ISM is used in this thesis for the investigation of realistic impulse responses of rooms. Pole-zero models describing the RIR mathematically are introduced in sect. §4.4 and are used in this dissertation for modelling of the acoustic channel, particularly in Chaps. 6 and 9. Noise models are discussed in sect. §4.6 and are used in Chap. 6. The discussion in this chapter is summarised in sect. §4.7.

4.2 The room transfer function

The acoustic response in an enclosed space between a sound source and a receiver is the result of the direct-path signal and all its reflections. Thus the sound wave can be modelled by the superposition of all sound waves in the room. Similar to the description of the propagation of the airflow from the lungs through the vocal tract in sect. §3.2 using Webster's equation in eqn. (3.1) on page 33, the propagation of sound in rooms can be described in terms of the propagation of pressure waves. Assuming that the sound wave travels through a homogenous medium at rest and independent of the wave amplitude, the following partial differential equation (PDE) describes the propagation of the sound pressure in terms of the observer position:

$$\frac{1}{c^2}\frac{\partial^2 p_{\mathbf{r}_o}(t)}{\partial t^2} - \nabla^2 p_{\mathbf{r}_o}(t) = s_{\mathbf{r}_s}(t), \tag{4.1}$$

where $p_{\mathbf{r}_o}(t)$ is the sound pressure at time t and observed at position $\mathbf{r}_o = (x, y, z)$ in Cartesian coordinates, $s_{\mathbf{r}_s}(t)$ is a sound excitation as a function of the source position, $\mathbf{r}_s$, and time, t, and $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ denotes the Laplacian operator. According to eqn. (4.1), if the power of the sound source is doubled, the sound pressure at some distant point doubles as well [154]. In practice, air is not completely at rest due to temperature variations, leading to nonlinearities in eqn. (4.1). Nonetheless, the effects due to inhomogeneity of air in the room are small and negligible for modelling purposes of a general expression of the acoustical properties of a room.

In sect. §3.3, the transfer function of the vocal tract was derived in order to obtain a relation between the input and output of the acoustic system. In the same way, it is desirable to derive the transfer function of the room, i.e.,

$$H_{(\mathbf{r}_o,\mathbf{r}_s)}(\omega) = \frac{S_{\mathbf{r}_s}(\omega)}{P_{\mathbf{r}_o}(\omega)}. \tag{4.2}$$

where $H_{(\mathbf{r}_o,\mathbf{r}_s)}(\omega)$ is the RTF, dependent on the observed and source positions, $\mathbf{r}_o$ and $\mathbf{r}_s$ respectively, and $P_{\mathbf{r}_o}(\omega)$ and $S_{\mathbf{r}_s}(\omega)$ are the Fourier transforms of $p_{\mathbf{r}_o}(t)$ and $s_{\mathbf{r}_s}(t)$ respectively. Assuming a closed, rectangular room, the wave equation in eqn. (4.1) can be solved and the RTF in eqn. (4.2) derived as a function of the resonant frequencies of the room (see Appendix A.6) [4]:

$$H_{(\mathbf{r}_o,\mathbf{r}_s)}(\omega) = G \sum_{i=0}^{\infty} \frac{P_i(\mathbf{r}_o)\, P_i(\mathbf{r}_s) \jmath\omega}{\omega^2 - \omega_i^2 - 2\jmath\delta_i\, \omega_i} \tag{4.3}$$

where $P_i(\mathbf{r}_s)$ and $P_i(\mathbf{r}_o)$ are the mutually orthogonal eigenfunctions of the resonant frequencies, ω_i, ω is the angular frequency, G is a gain factor, and δ_i is the damping constant according to the quality factor.

The expression of the RTF in eqn. (4.3) describes the acoustic properties of a reverberant room and can be used to model the reverberant channel. The following two sections discuss how the solution to the acoustic wave equation can be approximated by either simulating the impulse response of the room (sect. §4.3) or develop mathematical models based on the expression of the transfer function (sect. §4.4).

4.3 Simulating room acoustics

The image-source method proposed in [155] *simulates* the acoustic impulse response of a reverberant geometric room by considering that a reflected sound wave can be represented by the direct-path signal of an image source that is the mirror image by the reflecting wall.

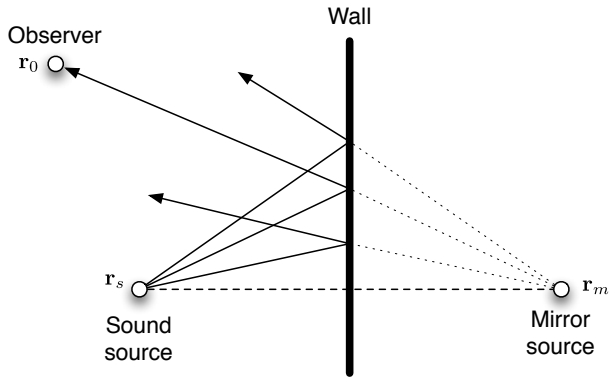


Figure 4.3: *Construction of a mirror source*

A sound wave is usually reflected from rigid surfaces with an angle equal to the incoming angle of the wave according to Snell's law. Assuming a point source in front a wall, each reflection from the wall of the sound wave can thus be thought of as originating from a virtual source (called image source) behind the wall emitting the same source signal as the original source and located on a perpendicular line to the wall and at the same distance as the original point source as illustrated in Fig. 4.3 [4]. If the power spectrum of the image source is adjusted to account for the absorption coefficient of the wall, the effect of the wall is replaced by the image source.

Assuming both the source and image source are single frequency point sources of acceleration in free space, their pressure waves can be expressed as [155]:

$$H_{(\mathbf{r},\mathbf{r}_o)}(\omega) = \frac{\exp\left[\jmath\omega\left(\frac{r}{c} - t\right)\right]}{4\pi r}$$

where $\mathbf{r} = \begin{bmatrix} x & y & z \end{bmatrix}^T$ denotes the position of some point source, and $r = |\mathbf{r} - \mathbf{r}_o|$ is the distance between the point source and the sensor. Hence, the pressure wave, $\hat{H}_{(\mathbf{r}_s,\mathbf{r}_o)}(\omega)$, at the receiver consists of the pressure wave from the source, $H_{(\mathbf{r}_s,\mathbf{r}_o)}(\omega)$, and that of the image source, $H_{(\mathbf{r}_m,\mathbf{r}_o)}(\omega)$ at position $\mathbf{r}_i = \begin{bmatrix} x_i & y_i & z_i \end{bmatrix}^T$, i.e.,

$$\hat{H}_{(\mathbf{r}_s,\mathbf{r}_o)}(\omega) = H_{(\mathbf{r}_s,\mathbf{r}_o)}(\omega) + H_{(\mathbf{r}_i,\mathbf{r}_o)}(\omega) = \left[\frac{\exp\left(\jmath\omega\frac{r_s}{c}\right)}{4\pi r_s} + \frac{\exp\left(\jmath\omega\frac{r_m}{c}\right)}{4\pi r_m}\right]\exp\left(-\jmath\omega t\right)$$

Note that if the wall is set at $x = 0$, $r_s^2 = (x_s - x_o)^2 + (y_s - y_o)^2 + (z_s - z_o)^2$ and $r_m = (x_s + x_o)^2 + (y_s - y_o)^2 + (z_s - z_o)^2$. This is due to the fact that the image source is placed at the same distance as the sound source at the opposite side of the wall.

For the case where the source is enclosed in a room, i.e., by six walls, each image itself is imaged, such that the pressure wave observed at the receiver is the superposition of all sources, i.e., [155–157]

$$\hat{H}_{(\mathbf{r}_s,\mathbf{r}_0)}(\omega) = \exp(-\jmath\omega t) \sum_{p=1}^{8} \sum_{q=-\infty}^{\infty} \frac{\exp\left(\jmath\omega \frac{\mathbf{r}_p + \mathbf{r}_q}{c}\right)}{4\pi |\mathbf{r}_p + \mathbf{r}_q|} \tag{4.4}$$

where the permutations of $\mathbf{r}_p = \begin{bmatrix} x_s \pm x_o & y_s \pm y_o & z_s \pm z_o \end{bmatrix}^T$ are given by $\{\mathbf{r}_p\}_{p=1}^{8}$, and $\mathbf{r}_q = 2\begin{bmatrix} n L_x & \ell L_y & m L_z \end{bmatrix}^T$ for a room of dimensions $L_x \times L_y \times L_z$ (width $\times$ length $\times$ height). Taking the inverse Fourier transform (IFT) of eqn. (4.4), the RIR is

$$h_{(\mathbf{r}_s,\mathbf{r}_0)}(t) = \sum_{p=1}^{8} \sum_{q=-\infty}^{\infty} \frac{\delta(t - \frac{|\mathbf{r}_p - \mathbf{r}_q|}{c})}{|\mathbf{r}_p - \mathbf{r}_q|} \tag{4.5}$$

which is the exact solution of the wave equation in a rectangular, rigid-walled (i.e., lossless) room [155]. By explicitly writing out the leftmost summation in eqn. (4.5) the RIR can be expressed in a more general form as

$$h_{(\mathbf{r}_s,\mathbf{r}_0)}(t) = \sum_{q=-\infty}^{\infty} \mathbf{g}_q^T \boldsymbol{\delta}(t - \boldsymbol{\tau}_q), \tag{4.6}$$

where $\mathbf{g}_q = \begin{bmatrix} \frac{1}{|\mathbf{r}_1 - \mathbf{r}_q|} & \dots & \frac{1}{|\mathbf{r}_8 - \mathbf{r}_q|} \end{bmatrix}^T$, $\boldsymbol{\tau}_q \triangleq \frac{1}{c}\begin{bmatrix} |\mathbf{r}_1 - \mathbf{r}_q| & \dots & |\mathbf{r}_8 - \mathbf{r}_q| \end{bmatrix}^T$, and $\boldsymbol{\delta}(t - \boldsymbol{\tau}_q)$ is a 8×1 vector of Dirac-delta functions, where each row is shifted by the corresponding $\{\tau_{p,q}\}_{p=1}^{8}$. This formulation leads to two important conclusions:

- As the RIR (but also the RTF in eqn. (4.4)) is expressed in terms of the source position, $\mathbf{r}_s$ and the sensor position, $\mathbf{r}_o$, the RIR (and RTF) varies with changing source-sensor positions and distances [158, 159]. As the source-sensor position varies with time, the RIR and RTF vary with time.
- Eqn. (4.6) suggests that the RIR can be modelled using a linear time-varying (LTV) representation. If $\mathbf{r}_s$ and $\mathbf{r}_o$ are fixed, the RIR is invariant and can be modelled using linear time-invariant (LTI) representations. Therefore, as eqn. (4.5) and hence eqn. (4.6) are the solution to the wave equation for rectangular rooms with rigid walls, the wave equation can be modelled using pole-zero models.

Sect. §4.4 therefore discusses pole-zero modelling of the acoustic channel. As the presence of zeros in the model generally leads to computational overhead, static RIRs are often approximated by all-pole models as discussed in sect. §4.4.1. Model order selection of the all-pole model is discussed in sect. §4.4.2.

4.4 Pole-zero modelling of room transfer functions

By assuming a rectangular room with perfectly rigid walls, and ignoring practical aspects such as temperature changes, opening and closing of doors and windows, or people and objects within the room, the solution of the wave equation is of the form of a rational function. This result suggests that the RTF can be approximated by a pole-zero model of the form [160]

$$H^{PZ}_{(\mathbf{r}_s,\mathbf{r}_o)}(\omega) = G^{PZ}_{(\mathbf{r}_s,\mathbf{r}_o)}\, z^R \frac{\prod\limits_{k\in\mathcal{Q}} \left(1 - q_k^{PZ} z^{-1}\right)}{\prod\limits_{i\in\mathcal{P}} \left(1 - p_i^{PZ} z^{-1}\right)} = \frac{\sum\limits_{k\in\mathcal{Q}+\mathcal{R}} d_k^{PZ} z^{-k}}{1 + \sum\limits_{i\in\mathcal{P}} b_i^{PZ} z^{-k}} \tag{4.7}$$

where P is the number of poles, $Q+R$ is the total number of zeros including those at the origin, $G^{PZ}(\mathbf{r}_s, \mathbf{r}_o)$ is a gain constant, $\left\{p_i^{PZ}\right\}_{i\in\mathcal{P}}$ are the poles, $\left\{q_i^{PZ}\right\}_{k\in\mathcal{Q}}$ are the zeros, $\left\{b_i^{PZ}\right\}_{i\in\mathcal{P}}$ are the autoregressive (AR) coefficients and $\left\{d_k^{PZ}\right\}_{k\in\mathcal{Q}+\mathcal{R}}$ are the moving average (MA) coefficients. As the RTF is stable and causal, the denominator must be stable and causal and hence the poles lie within the unit circle, i.e., $|p_i^{PZ}| < 1 \,\forall\, i \in \mathcal{P}$. As RTFs are, however, often non-minimum phase, the zeros, q_k^{PZ} may lie outside of the unit circle. One can therefore distinguish between a minimum and non-minimum phase component of the zeros and rewrite eqn. (4.7) as

$$H^{PZ}_{(\mathbf{r}_s,\mathbf{r}_o)}(\omega) = G^{PZ}(\mathbf{r}_s, \mathbf{r}_o)\, z^R \frac{\prod\limits_{k\in\mathcal{Q}_m} \left(1 - r_k^{PZ} z^{-1}\right) \prod\limits_{\ell\in\mathcal{Q}_n} \left(1 - s_\ell^{PZ} z\right)}{\prod\limits_{i\in\mathcal{P}} \left(1 - p_i^{PZ} z^{-1}\right)} \tag{4.8}$$

where $\left\{r_k^{PZ}\right\}_{k\in\mathcal{Q}_m}$ lie within the unit circle and are the Q_m minimum phase components, whilst $\left\{s_\ell^{PZ}\right\}_{\ell\in\mathcal{Q}_n}$ lie outside the unit circle and are the Q_n non-minimum phase components.

As discussed in [159–161] acoustic impulse responses (AIRs) are, in general, very long. The inclusion of zeros in eqn. (4.8) thus typically requires $n_s = T_{60} f_s$ coefficients only to model the zeros in the model. For example, if $T_{60} = 0.5$ seconds and $f_s = 10$ kHz, $n_s = 5000$ all-zero coefficients are necessary. Furthermore, the all-zero part of eqn. (4.8) leads to large variations in the RTF for even small changes in source-observer positions such that the resulting pole-zero model may thus only be effective for limited spatial combinations of source and receiver positions [159, 160, 162, 163]. Instead, the pole-zero model is often reduced to all-pole models to approximate the RTF.

4.4.1 All-pole model of the RTF

All-pole models are popular methods for approximating transfer functions, such as that of the human vocal tract, as discussed in sect. §3.3.2 on page 41. The poles of the model describe the resonances of the standing waves in a room, allowing for satisfactory representation of the RTF and a reduction in complexity as compared to pole-zero or all-zero models [160]. All-pole models are also significantly less sensitive to changes in the source-sensor positions. As all-pole models are minimum phase (i.e., stable and causal), however, they cannot model the non-minimum phase components of RTFs. Nonetheless, subband all-pole models as proposed in [164, 165] can be used to overcome this shortcoming.

Using all-pole models,the RTF can be expressed of the form:

$$H_{(\mathbf{r}_s,\mathbf{r}_o)}(\omega)^{AP} = \frac{G^{AP}(\mathbf{r}_s,\mathbf{r}_o)}{\prod_{i\in\mathcal{P}}\left(1-p_i^{AP}z^{-1}\right)} = \frac{G^{AP}(\mathbf{r}_s,\mathbf{r}_o)}{1+\sum_{i\in\mathcal{P}} b_i^{AP}z^{-i}}. \tag{4.9}$$

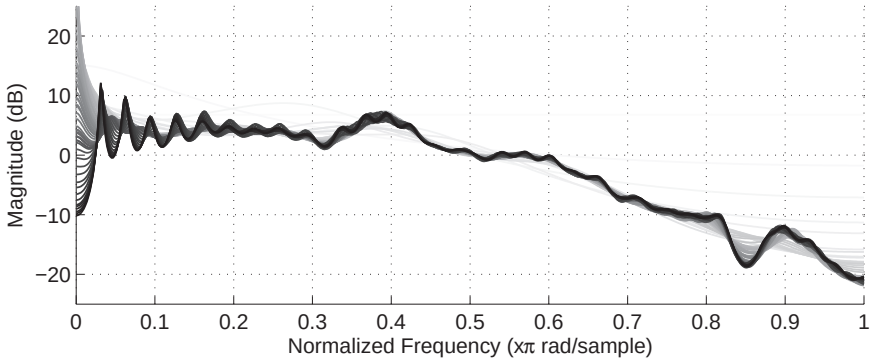
where $\left\{p_i^{AP}\right\}_{i\in\mathcal{P}}$ are the poles of the model and $\left\{b_i^{AP}\right\}_{i\in\mathcal{P}}$ are the corresponding AR coefficients. Speech signals distorted by a reverberant channel can thus be modelled as

$$y_t = \sum_{p\in\mathcal{P}} b_p\, y_{t-p} + x_t \tag{4.10}$$

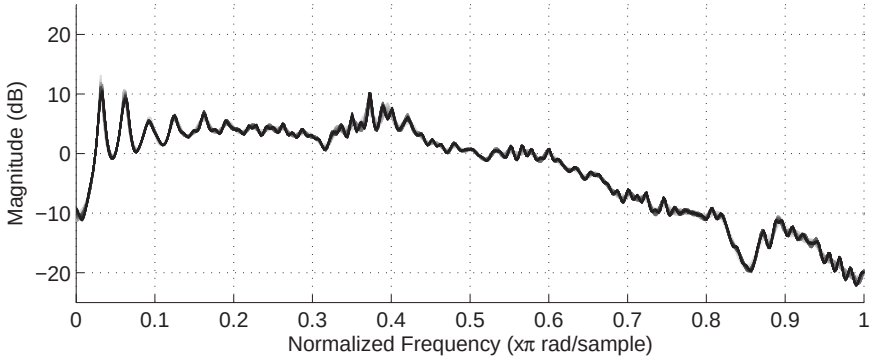
where x_t is the source signal as before, y_t is the distorted observed signal, and $\{b_p\}_{p\in\mathcal{P}}$ are the P channel parameters.

The all-pole approximation of the response of an acoustic horn, modelled by an all-pole filter using a model-order of $P = 72$ is illustrated in Fig. 4.5. The model-order was chosen according to the results in [69], where $P = 72$ was found to be the optimum model order for the acoustic response. The corresponding pole positions are shown in Fig. 4.5b. A comparison of the frequency response of the acoustic horn with the response of its all-pole model is shown in Fig. 4.5a and illustrates the satisfactory representation of the resonant peaks. The all-pole model for approximation of static RTFs is therefore used in Chap. 6 to Chap. 10 for the dereverberation of speech from stationary speakers.

In practice, the model orders are generally unknown, and *estimates* of the channel order, P_{est}, have to be used. If more model parameters than in the actual model are used – i.e., $P_{est} > P$ – the model is *over-modelled*. Usage of less parameters – i.e., $P_{est} < P$ – is referred to as *under-modelling*. To illustrate the effects of over- and under-



(a) *Acoustic horn modelled at model orders varying between* 1 *(grey) and* 72 *(black)*



(b) *Acoustic horn modelled at model order varying between* 72 *(grey) and* 200 *(black)*

Figure 4.4: *Modelling of horn acoustic response using an increasing channel order; Optimal order at* $\mathrm{P} = 72$.

modelling, the frequency responses for $\mathrm{P}_{\mathrm{est}} \leq 72$ are shown in Fig. 4.4a, and $\mathrm{P}_{\mathrm{est}} \geq 72$ are shown in Fig. 4.4b. As seen from these figures, under-modelling causes the omission of some resonant peaks, whereas over-modelling introduces additional resonant peaks in the spectrum. Thus, under-modelling can lead to the omission of high-energy taps required for the description of the acoustic channel. In contrast, over-modelling can result in the modelling of artificial characteristics of the channel.

In cases where the frequency response of the channel is known, it is desirable to extract the model order, corresponding to the number of resonant peaks. The estimation of model orders from known frequency responses is pertinent in Chap. 10, where the complexity of the proposed dereverberation framework is derived and shown to grow quadratically with the channel model order. In order to reduce the computational overhead, a multirate extension to dereverberation framework is proposed, segmenting the fullband channel into several subbands in order to reduce the model order in each frequency band. To show the computational benefit of multirate processing, the model orders of a fullband and subband channel are investigated. Methods for the estimation of model orders are hence necessary.

4.4.2 Theoretical pole order

Recalling that the all-pole transfer function in eqn. (4.9) is an approximation to the solution of the acoustic wave equation in eqn. (4.1), the order of the all-pole model is related to the number of modes in the characteristic equation (see Appendix A.6). The number of modes, $N(f_u)$, for a room of dimensions $L_x \times L_y \times L_z$ with upper frequency f_u, is given by [4, 166]:

$$N(f_u) = \frac{4\pi}{3} V \left(\frac{f_u}{c}\right)^3 + \frac{\pi}{4} S \left(\frac{f_u}{c}\right)^2 + \frac{L}{8} \left(\frac{f_u}{c}\right) \tag{4.11}$$

where $V = L_x L_y L_z$ is the volume of the room, $S = 2(L_x L_y + L_y L_z + L_x L_z)$ is the room surface area, and $L = 4(L_x + L_y + L_z)$ is the sum of the edge lengths occurring in the rectangular room. If $f_u < 500$Hz and $V \gg S$, the last two terms on the right hand side of eqn. (4.11) can be ignored, i.e., $\hat{N}(f_u) = \frac{4\pi}{3} V \left(\frac{f_u}{c}\right)^3$. This is often the case for large concert halls. For instance, the Tokyo Opera City Concert Hall built in 1997 and seating 1632 spectators is built as a shoebox style with a vaulting pyramid and is of size $20 \times 41.4 \times 27.6$m (width $\times$ depth $\times$ height) [167]. Its volume is therefore $2.29 \cdot 10^4 \text{m}^3$, whereas the surface area amounts to $5.05 \cdot 10^3 \text{m}^2$, such that $V \gg S$. Assuming an upper frequency of $f_u = 5$kHz, the error between $N(f_u)$ in eqn. (4.11) and $\hat{N}(f_u)$ is 0.28%, i.e., negligible.

Using $\hat{N}(f_u)$, the order of the all-pole model up to sampling frequency, f_s, is therefore given by [4, 166]

$$P(f_s) \approx 2N\left(\frac{f_s}{2}\right) = \frac{V\pi}{3}\left(\frac{f_s}{3}\right)^3. \tag{4.12}$$

For all-pole orders equal to $P(f_s)$, the all-pole model approximates the actual room response. For lower model orders, the poles represent major resonant frequencies of high Q-factors [4]. Recalling that even all-zero models require $n_s = T_{60} f_s$ parameters to approximate the RTF and comparing to eqn. (4.12), $P(f_s)$ requires more than n_s poles and hence represents a very loose upper bound [166]. Furthermore, the assumption that the volume of the room is significantly larger than the surface area only holds for large halls. For a standard-size shoebox office of the size $2.78 \times 4.68 \times 3.2$, for example, $S > V$ such that the underlying assumption of simplifying eqn. (4.11) to $\hat{N}(f_u)$ does not hold.

A more reliable model order estimate can be obtained by modelling the AIR as an all-pole filter by exciting the response with white Gaussian noise (WGN) and estimating the infinite impulse response (IIR) parameters using the Yule-Walker equations for

each choice of model order, $P = P_{min}, \ldots, P_{max}$, between a minimum and maximum model order. For each resulting set of AR parameters, the mean squared error (MSE), MSE_P, is calculated model order selection criteria such as the Akaike's information criterion (AIC) [168, Chapter 1.6], [169]:

$$AIC_P = 2P + N \ln MSE_P \tag{4.13}$$

or the minimum description length (MDL) [170, Chapter 9.3]:

$$MDL_P = P \ln N + N \ln MSE_P \tag{4.14}$$

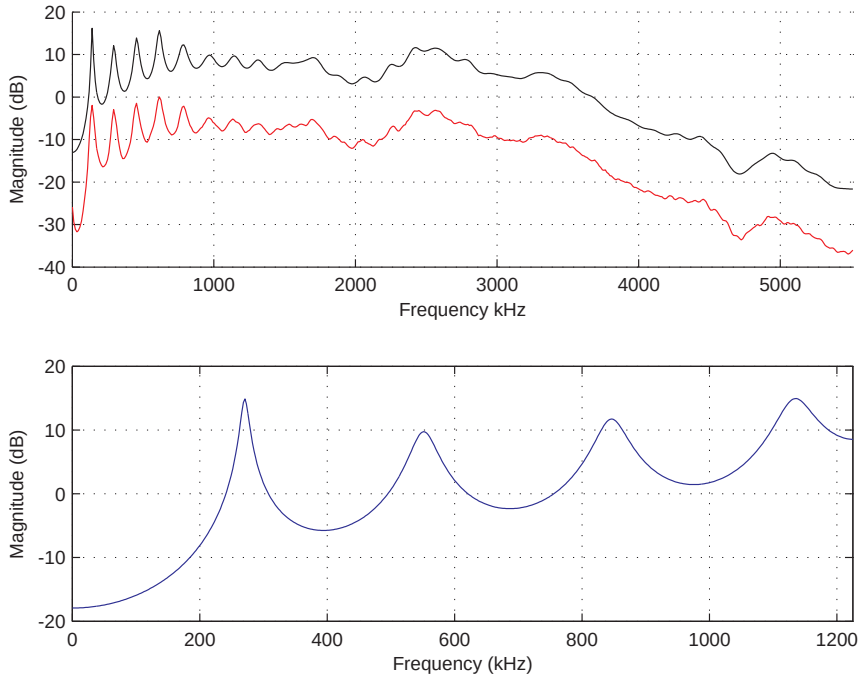
are applied, where N is the length of the number of samples. The optimal model order corresponds to the minimum model order selection criterion. As any restricting assumptions about the room size are avoided in eqns. (4.13) and (4.14), the AIC and MDL are used for model order selection in Chap. 10.

4.5 All-pole model of a gramophone horn response

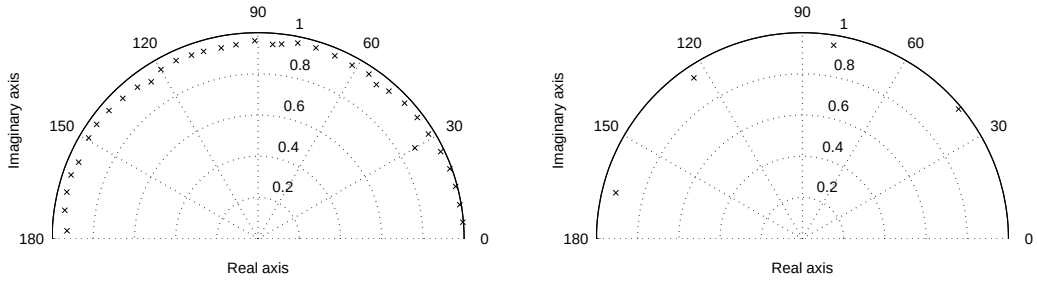
Instead of radiating a speech signal within a constrained room, consider a speech signal recorded through a gramophone horn. Due to reflections of the anechoic speech signal off the enclosing horn, the distorted speech signal exhibits a megaphonic quality. A typical gramophone horn is investigated in detail in Spencer and Rayner [68] and Spencer [69]. The response of the horn at a sampling rate of $f_2 = 11.025$kHz is shown in Fig. 4.5a. Whilst the horn exhibits a relatively flat high-frequency response, the response is resonant at low frequency.

As discussed in [69], the response can be approximated by an all-pole filter with optimal model order of $P = 72$. The magnitude response of the all-pole model is shown as a grey line in Fig. 4.5a, approximating the shape of the horn response well with an offset of approximately 20dB. The corresponding pole positions are shown in Fig. 4.5b. Due to the flatness of the response, the low-frequency response can be accurately modelled by an all-pole filter of model order $P = 8$ with sampling frequency $f_s = 2.45$kHz [45] as illustrated in the lower subplot of Fig. 4.5a. The resulting channel poles are displayed in Fig. 4.5c.

As the distortion of the anechoic speech signals due radiation through the gramophone horn can be considered as reverberation in a simplified enclosing environment, and the gramophone horn response in Fig. 4.5a can be accurately modelled using an all-pole filter of known model order, the experiments performed in this chapter are based on anechoic signals filtered the gramophone horn response.



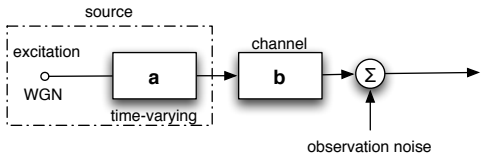
(a) *Frequency response of acoustic horn in [69] (red line) vs. AR(72) approximation (black line) vs. AR(8) approximation (blue line).*



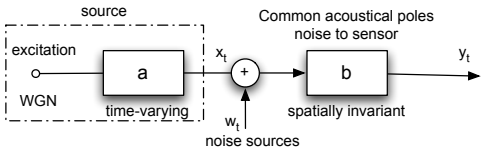
(b) *Position of AR poles for channel order 72*

(c) *Position of the AR poles for channel order 8*

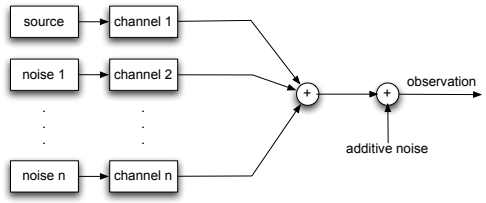
Figure 4.5: *Properties of acoustic horn channel*



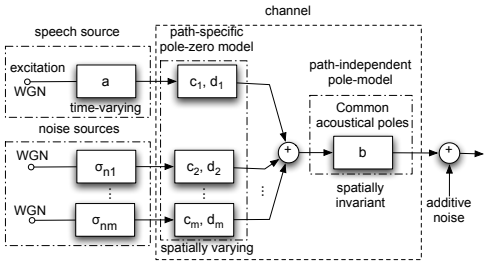
(a) *Conventional model assuming that observation noise is property of room sensed directly in front of receiver*



(b) *Model resulting from simplification of Fig. 4.6c to audio and noise sources filtered through the same channel*



(c) *Model considering distant noise sources filtered through separate channels*



(d) *Proposed model with spatially separated noise sources and separation of common and non-stationary acoustic poles and zeros*

Figure 4.6: *Modelling assumptions leading to proposed model in Fig. 4.6d*

In addition to the distorting effects of reverberation, a source signal exhibited in confined spaces is distorted by noise radiating from, e.g., computer fans. *Noise* models are therefore crucial components in the system model to describe interfering noise sources in the room.

4.6 Noise models

Noise models are often modelled as an additive common signal observed within the room, independent of the microphone's location and unaffected by acoustics of the room. Therefore, noise is modelled to be added at the output of the system (Fig. 4.6a). In a more realistic model, spatially distinct noise sources should be observed after they have propagated through the acoustic system, and therefore have corresponding but distinct AIRs (Fig. 4.2). Hence, a combination of signals filtered by separate channels is observed at the receiver (Fig. 4.6c).

However, due to the number of different RIRs and unknown variables that need to be identified in the system model, issues arise with estimating all the relevant system parameters in the model in Fig. 4.6c. The model can thus be simplified using the notion of common acoustical poles. In [163], Haneda *et al.* decompose individual channels into a combination of two components: one that is dependent on the source-sensor

geometry, and the other which is acoustically common to all source-sensor arrangements.

Motivated by the presence of common resonances, we propose to apply this separation to the model in Fig. 4.6c to obtain a more realistic room acoustical model as shown in Fig. 4.6d. Although the general model in Fig. 4.6d is of great interest, the presence of general RTFs dependent on the source-sensor geometry leads to difficulties in uniquely identifying the source signals in the blind deconvolution problem. Research into the, e.g., the identifiability of model parameters is before this model can be used with confidence. As this research exceeds the scope of this dissertation, a simplified model of Fig. 4.6c is in this thesis. It is assumed that the noise source is located close to the speakers, such that the path-specific channels between the speaker and sensor and the noise and sensor are identical. Hence, by rearrangement of Fig. 4.6c, the path-specific channels can be included in the common acoustical model, leading to the model in Fig. 4.6b. Eqn. (4.10) can thus be extended to

$$y_t = \sum_{p \in \mathcal{P}} b_p \, y_{t-p} + x_t + \sigma_{w_t} \, w_t \tag{4.15}$$

where $w_t \sim \mathcal{N}(0, 1)$ for WGN source with measurement noise variance $\sigma^2_{w_t}$.

4.7 Summary

This chapter discussed the room transfer function, describing the acoustic properties of a room, and its models. The RTF was derived from the acoustic wave equation. It was shown that the RTF of a room can be simulated using the ISM by modelling reflections off surrounding walls in a geometric room as additional sound sources outside of the room. The ISM is particularly helpful when measured RTFs are not available or when idealistic RTFs are desired. The ISM is used throughout this thesis to generate realistic room responses. In particular, Chap. 10 utilise the ISM for investigation of typical channel model orders for evaluation of the computational complexity of the proposed dereverberation approach.

Furthermore, it was shown that the RIR is of the form of a rational function and can therefore be modelled using pole-zero models. Due to the required number of zeros in the resulting model, pole-zero and all-zero models can lead to computational overhead. All-pole models are therefore often used instead to approximate the RIR. The derivation of the proposed dereverberation approach in Chap. 6 is hence based on the all-pole channel model. The required model order can be extracted from the magnitude response of the room as the number of resonant peaks, and can be estimated

using model selection criteria such as the AIC or MDL. Model order estimation is of particular interest in Chap. 10, where the computational complexity of the proposed dereverberation approach is evaluated and found to increase quadratically with the channel model order. Typical channel model orders are therefore investigated and a multi-rate extension is proposed to segment the full channel response into several sub-band responses of lower model order.

As, in practice, speech is also subject to interference by close-by noise sources, a noise model was proposed whereby the acoustic channel is separated into a path-specific channel and a channel common to the room at any position therein. However, research into the parameter identifiability is necessary before confident application in speech processing applications. Thus, a simplified model assuming one noise source close to the speaker was introduced instead.

Given the discussion of the source and channel model in Chaps. 2 and 3, the following chapter reviews the estimation approaches required in this thesis, concluding the background study and hence Part II.

Chapter 5

Bayesian estimation and sequential Monte Carlo methods

5.1 Introduction

Blind speech dereverberation approaches attempt to recover the clean signal from the reverberant observations. Due to the underlying assumptions of many approaches in the literature, however, the solution to blind speech dereverberation is restricted to only small subgroups of the problem as discussed in Chap. 2.

As discussed in Chap. 1, the aim of this thesis is to investigate whether blind speech dereverberation techniques can be improved upon and a flexible and extendible framework be developed by considering the problem from a Bayesian perspective. In the Bayesian sense, the problem of blind speech dereverberation is viewed as the selection of the best possible representation of the source signal out of all potential candidates using the knowledge of all past observations. In other words, the posterior probability density functions (pdfs) of the set of desired variables (in this case the source signal as well as the source and channel parameters) are constructed and point estimates of the unknown variables are drawn from the posterior pdfs by either optimisation, e.g., using maximum a posteriori (MAP) estimates, or by the formation of expectations, e.g., using minimum mean-square error (MMSE) estimates.

In order to construct the posterior pdf, i.e., to infer knowledge from the distorted observations, Bayes's theorem is exploited. Bayes's theorem naturally relates the *known* observations, $\mathbf{y}_{1:t}$, and the *unknown* variables, $\boldsymbol{\varphi}_{0:t}$, via

$$\mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}, \mathcal{M}\right) = \frac{\mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}, \mathcal{M}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathcal{M}\right)}{\mathrm{p}\left(\mathbf{y}_{1:t} \mid \mathcal{M}\right)}. \tag{5.1}$$

where $\boldsymbol{\varphi}_{0:t} = \begin{bmatrix} \boldsymbol{\varphi}_0 & \dots & \boldsymbol{\varphi}_t \end{bmatrix}^T$ is the collection of unknown parameters up to time t, where $\boldsymbol{\varphi}_t = \begin{bmatrix} \mathbf{x}_t & \theta_t & \mathbf{b} \end{bmatrix}^T$ contains the collection of the Q most current source signal samples $\mathbf{x}_t = \begin{bmatrix} x_t & \dots & x_{t-Q+1} \end{bmatrix}^T$, the channel parameters, $\mathbf{b} = \begin{bmatrix} b_1 & \dots & b_P \end{bmatrix}^T$ as well as the source model parameters and noise variance terms, $\theta_t = \begin{bmatrix} \mathbf{a}_t & \phi_{\nu_t} & \boldsymbol{\phi}_{\mathbf{w}_t}^T \end{bmatrix}^T$. Furthermore, $\mathbf{y}_{1:t} = \begin{bmatrix} y_0 & \dots & y_t \end{bmatrix}^T$ are the observations available up to time t, $\mathcal{M}$ is the underlying model, $p(\boldsymbol{\varphi}_{0:t} | \mathbf{y}_{1:t}, \mathcal{M})$ is the posterior pdf, $p(\mathbf{y}_{1:t} | \boldsymbol{\varphi}_{0:t}, \mathcal{M})$ is the likelihood function, $p(\boldsymbol{\varphi}_{0:t} | \mathcal{M})$ is the prior pdf and $p(\mathbf{y}_{1:t} | \mathcal{M})$ is the evidence term, given by

$$p(\mathbf{y}_{1:t} | \mathcal{M}) = \int_{\mathbb{R}^{\varphi}} p(\mathbf{y}_{1:t} | \boldsymbol{\varphi}_{0:t}, \mathcal{M}) \, p(\boldsymbol{\varphi}_{0:t} | \mathcal{M}) \, \mathrm{d}\boldsymbol{\varphi}_{0:t}, \tag{5.2}$$

where $\int_{\Phi}(\cdot) \, \mathrm{d}\boldsymbol{\varphi}_{0:t} \triangleq \int_{\Phi} \dots \int_{\phi}(\cdot) \, \mathrm{d}\boldsymbol{\varphi}_1 \dots \mathrm{d}\boldsymbol{\varphi}_t$ over the region of support Φ of $\boldsymbol{\varphi}_{0:t}$. Note that the model, $\mathcal{M}$ is constant and is therefore ignored in the following for brevity.

Bayesian inference thus allows the specification of the probability of an unknown variable and, hence, makes probabilistic statements directly about the unknown parameters. As Bayesian inference is closely tied to prior knowledge, its estimation performance is highly dependent on the choice of underlying system models. Bearing in mind that the vocal tract mechanism as well as room acoustics have been thoroughly researched for decades, a multitude of accurate models exist for both the speech production system as well as the room transfer function. Bayesian inference thus seems particularly appealing for blind speech dereverberation problems.

This chapter motivates the idea of Bayesian statistics and introduces the methods required in Part III of this thesis. Point estimation from the posterior pdf using maximum-likelihood (ML), MAP, and MMSE estimates is discussed in sect. §5.2 and will be used throughout Part III. The Kalman filter is the optimal estimator of linear state spaces in the MMSE sense and is discussed, including its non-linear variants, in sect. §5.3. Kalman filtering is utilised in Part III for estimation of the source signal and the autoregressive (AR) parameters of the room impulse response (RIR) and is extensively used in Chap. 6. An underlying assumption of the Kalman filter is that the model parameters of the system are known. In blind speech dereverberation problems, however, only the received and distorted observations are available.

For unknown system parameters, an ensemble of Kalman filters could theoretically be evaluated for every possible parameter choice and selecting the solution with the highest likelihood. To avoid the involved dimensionality issues and computa-

tional overhead, an ensemble of Kalman filters could be evaluated for *stochastically* selected parameters.

This concept directly leads to Monte Carlo methods, a class of algorithms that repeatedly draw random samples to obtain their results as discussed in sect. §5.4. Sequential Monte Carlo (SMC) methods sequentially sample a large cloud / set of random variates, allowing for real-time processing (see sect. §5.5). Kalman filters have previously been integrated in SMC frameworks for the estimation of analytically tractable substructures, commonly known as Rao-Blackwellized SMC (see sect. §5.6). Rao-Blackwellized SMC forms the crucial basis for the blind speech dereverberation algorithm proposed in Part III. The discussion and key points of this chapter are summarised in sect. §5.7.

5.2 Types of estimators

Based on the speech production and room acoustical model developed in Chaps. 3 and 4, estimators of the clean speech signal and unknown system model parameters should be developed. Estimators that exhibit optimality with respect to certain criteria are extremely appealing, as the best possible estimation performance with respect to the chosen criterion is guaranteed. Popular criteria are the maximisation of the likelihood function of the measured data, or the posterior pdf, as well as the minimisation of the mean squared error (MSE).

5.2.1 Maximum likelihood estimators

If models of the source and channel are available in probabilistic form, statistical estimation theory can be used for signal enhancement. A simplistic choice is to use ML estimators that obtain point estimates the unknown variables, $\boldsymbol{\varphi}_{0:t}$, by maximising the likelihood of the distorted observations, $\mathbf{y}_{1:t}$, i.e.,

$$\boldsymbol{\varphi}_{0:t}^{\mathrm{ML}} = \arg\max_{\boldsymbol{\varphi}_{0:t}} \mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right). \tag{5.3}$$

ML estimation thus infers knowledge about the source signal, channel, and parameters from only the observations.

As only knowledge about the measured data is taken into account, ML estimators are based on purely objective observations. ML estimation is thus of particular interest for one-dimensional and bounded unknown variables, where maximisation of the likelihood in eqn. (5.3) needs to be performed for a single $\boldsymbol{\varphi}_t$ within a predetermined region of support. Nonetheless, as the dimension of $\boldsymbol{\varphi}_t$ increases and the variables

in $\boldsymbol{\varphi}_t$ are unbounded within the whole region of real numbers, the optimisation of eqn. (5.3) causes dimensionality and, hence, computational issues.

In order to reduce the dimension of the parameter space over which eqn. (5.3) is maximised, *subjective* information by means of prior pdfs can be incorporated.

5.2.2 Maximum *a posteriori* estimators

Prior belief and knowledge about the system is provided by the source and channel parameter models discussed in Chaps. 3 and 4 and can be exploited to improve estimation. Estimates of the unknown variables are obtained by construction of the posterior pdf of the states and parameters, $p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})$, which is related to the likelihood, $p(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t})$, via Bayes's theorem in eqn. (5.1). To obtain the optimal value of $\boldsymbol{\varphi}_{0:t}$, MAP estimates can be evaluated by maximisation with respect to the variables of interest, i.e.,

$$\boldsymbol{\varphi}_{0:t}^{\mathrm{MAP}} = \arg\max_{\boldsymbol{\varphi}_{0:t}} p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}) \tag{5.4}$$

As compared to ML estimators, MAP estimators thus incorporate knowledge about the distributions of the unknowns. Hence, uncertainty is introduced in the estimation of unknown parameters in order to update *belief* in the estimates as new data becomes available. MAP estimation is therefore sometimes interpreted as *penalised* ML [171]. As MAP estimators are based on the posterior pdf, related to the data via Bayes's theorem in eqn. (5.1), eqn. (5.4) is part of the group of Bayesian estimators.

5.2.3 Minimum mean squared error estimators

The MMSE estimate is given by minimising the MSE between the unknown variables, $\boldsymbol{\varphi}_{0:t}$, and an estimate, $\hat{\boldsymbol{\varphi}}_{0:t}$ and is expressed as

$$\boldsymbol{\varphi}_{0:t}^{\mathrm{MMSE}} = \mathbb{E}_{p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})}[\boldsymbol{\varphi}_{0:t}] \triangleq \int \boldsymbol{\varphi}_{0:t}\, p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})\, \mathrm{d}\boldsymbol{\varphi}_{0:t}. \tag{5.5}$$

For completeness, the derivation of eqn. (5.5) can be found in Appendix B.1. Similar to the MAP estimator, MMSE estimators are part of Bayesian estimators due to the utilisation of the posterior pdf in eqn. (5.5).

An MMSE estimator that is optimal for linear and Gaussian state spaces and extensively used in Chap. 6 is the Kalman filter as discussed in the following section.

5.3 MMSE estimation using the Kalman filter

The Kalman filter, developed by Kalman in [10], is a widely applied tool in statistical signal processing, aimed at sequential estimation of the unobserved states of a linear dynamic system. Estimates of the states are predicted based on *a priori* knowledge and corrected based on knowledge inferred from the distorted measurements. The Kalman filter is the optimal estimator of the states in the MMSE sense when the states and measurements obey a conditionally Gaussian state-space (CGSS) formulation, i.e., the posterior pdfs of both the states and measurements are Gaussian.

The underlying system model is assumed to be of the form:

$$\mathbf{x}_t = \mathbf{A}_t\,\mathbf{x}_{t-1} + \mathbf{\Sigma}_{\mathbf{v}_t}\,\mathbf{v}_t \qquad \mathbf{v}_t \sim \mathcal{N}\left(\mathbf{0}_{Q\times 1},\, \mathbf{I}_Q\right) \tag{5.6a}$$

$$\mathbf{y}_t = \mathbf{C}_t\,\mathbf{x}_t + \mathbf{\Sigma}_{w_t}\,\mathbf{w}_t \qquad \mathbf{w}_t \sim \mathcal{N}\left(\mathbf{0}_{M\times 1},\, \mathbf{I}_M\right) \tag{5.6b}$$

where $\mathbf{x}_t = \begin{bmatrix} x_t & \dots & x_{t-Q+1} \end{bmatrix}^T$ is vector of the past and current Q source signal samples (or hidden states of the state space), $\mathbf{A}_t$ is the (known) transition matrix from the state at $t-1$ to the state at t, $\mathbf{\Sigma}_{\mathbf{v}_t}$ is the covariance matrix of the excitation noise, $\mathbf{y}_t = \begin{bmatrix} y_{1,t} & \dots & y_{M,t} \end{bmatrix}^T$ is the observed signal sample at each of the M sensors, $\mathbf{C}_t$ is the (known) transformation matrix from the states to the observations, and $\mathbf{\Sigma}_{w_t}$ is the measurement noise covariance. It is important to note that the transition and transformation matrices, as well as the covariance terms of the excitation and measurement noise are *known*, such that estimation of the states is not blind.

Due to the Gaussian excitation and linear structure of the state space in eqn. (5.6), the posterior pdf of the source signal, $p\left(\mathbf{x}_t \mid \mathbf{y}_{1:t}\right)$, is normally distributed and fully specified by its mean and covariance. Since the maximum of a Gaussian is located at its mean, the mean of the posterior pdf corresponds to the MAP estimate of the source signal.

The states are predicted using the pdf $p\left(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}\right)$, obtained by marginalising the state, $\mathbf{x}_{t-1}$, at the previous time, $t-1$, from $p\left(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}, \mathbf{x}_{t-1}\right)$, i.e., [17]

$$\begin{aligned} p\left(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}\right) &= \int p\left(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}, \mathbf{x}_{t-1}\right) p\left(\mathbf{x}_{t-1} \mid \mathbf{y}_{1:t-1}\right) \mathrm{d}\mathbf{x}_{t-1} \\ &= \mathcal{N}\left(\mathbf{x}_t \,\middle|\, \boldsymbol{\mu}_{t|t-1},\, \mathbf{\Sigma}_{t|t-1}\right) \end{aligned} \tag{5.7}$$

where $p\left(\mathbf{x}_{t-1} \mid \mathbf{y}_{1:t-1}\right)$ is the marginal posterior pdf obtained at time $t-1$. Therefore, eqn. (5.7) outlines that rather than retaining the entire state trajectory, $\mathbf{x}_{0:t-1}$, it is sufficient to store the most recent marginal pdf, $p\left(\mathbf{x}_{t-1} \mid \mathbf{y}_{1:t-1}\right)$ only.

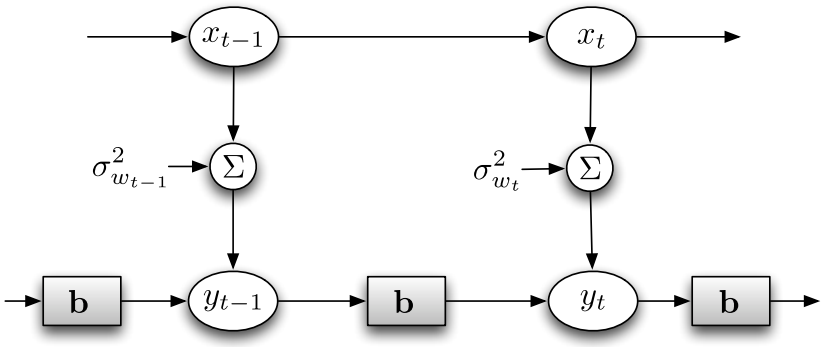


Figure 5.1: *Interpretation of the observations as measurements of a hidden Markov model*

Using the predicted posterior pdf in eqn. (5.7), the posterior pdf can be updated using Bayes's theorem via [17]

$$\mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t}\right) = \frac{\mathrm{p}\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \mathbf{x}_t\right) \mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}\right)}{\mathrm{p}\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}\right)} = \mathcal{N}\left(\mathbf{x}_t \,\middle|\, \boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}\right) \tag{5.8}$$

where $\mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}\right)$ is the pdf of the states predicted using the observations only, the evidence term, $\mathrm{p}\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}\right)$, can be obtained using eqn. (5.2) and is independent of $\mathbf{x}_t$, acting as a normalising scaling constant, and the likelihood function, $\mathrm{p}\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_t\right)$, can be obtained straightforwardly probability transformation to eqn. (5.6b), such that

$$\mathrm{p}\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{y}_t \,\middle|\, \mathbf{C}_t\, \mathbf{x}_t, \boldsymbol{\Sigma}_{w_t}\right). \tag{5.9}$$

The desired posterior pdf at time t therefore can be obtained by inferring knowledge from most recent observations at time t via $\mathrm{p}\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \mathbf{x}_t\right)$, to update the predicted pdf, $\mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t-1}\right)$.

By recursively inserting eqns. (5.9) and (5.7) into eqn. (5.8), the mean and covariance can be found as (see, e.g., [17, 172])

$$\boldsymbol{\mu}_{t|t-1} = \mathbf{A}_t \boldsymbol{\mu}_{t-1|t-1} \tag{5.10a}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \mathbf{A}_t \boldsymbol{\Sigma}_{t-1|t-1} \mathbf{A}_t^{\mathsf{T}} + \boldsymbol{\Sigma}_{\mathbf{v}_t}\, \boldsymbol{\Sigma}_{\mathbf{v}_t}^{\mathsf{T}} \tag{5.10b}$$

$$\boldsymbol{\mu}_{t|t} = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t \left(\mathbf{y}_t - \mathbf{C}_t\, \boldsymbol{\mu}_{t|t-1}\right) \tag{5.10c}$$

$$\boldsymbol{\Sigma}_{t|t} = \left(\mathbf{I}_M - \mathbf{K}_t \mathbf{C}_t\right) \boldsymbol{\Sigma}_{t|t-1} \tag{5.10d}$$

where

$$\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1} \mathbf{C}_t^{\mathsf{T}} \left(\mathbf{C}_t \boldsymbol{\Sigma}_{t|t-1} \mathbf{C}_t^{\mathsf{T}} + \boldsymbol{\Sigma}_{\mathbf{w}_t}\right)^{-1} \tag{5.11}$$

is a weighting matrix referred to as the Kalman gain. Eqs. (5.10) are known as the prediction (or propagation) and update equations of the Kalman filter, also referred to as the Kalman equations and can be interpreted as follows.

1. In the prediction (or propagation) step, the estimate, $\boldsymbol{\mu}_{t|t-1}$, of $\mathbf{x}_t$ is obtained based on all the previous observations, $\mathbf{y}_{1:t-1}$ and the estimate of the unknown variables, $\boldsymbol{\mu}_{t-1|t-1}$, at time $t-1$.
2. In the correction (or update) step, the predicted estimate, $\boldsymbol{\mu}_{t|t-1}$, is updated using the new information available through $\mathbf{y}_t$. $\mathbf{K}_t\left(\mathbf{y}_t - \mathbf{C}_t\,\boldsymbol{\mu}_{t|t-1}\right)$ can be considered as a weighting term of the additional information acquired through $\mathbf{y}_t$ in the estimate, $\boldsymbol{\mu}_{t|t}$, as compared to the predicted estimate, $\boldsymbol{\mu}_{t|t-1}$. If $\mathbf{y}_t$ contains no new information, the additive term is equal to zero (or contains no entropy or innovation) and $\boldsymbol{\mu}_{t|t-1}$ is "good enough" to reflect an estimate, $\boldsymbol{\mu}_{t|t}$, of $\mathbf{x}_t$. If $\mathbf{y}_t$ contains new information, the additive term is non-zero and $\mathbf{y}_t$ is incorporated when updating the estimates.

The covariance terms, $\boldsymbol{\Sigma}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t}$ in eqns. (5.10b) and (5.10d) are of particular interest for performance evaluation of the estimates: as shown in Appendix B.3, the diagonal terms in $\boldsymbol{\Sigma}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t}$ are equivalent to the MSE of the corresponding elements in $\boldsymbol{\mu}_{t|t-1}$ and $\boldsymbol{\mu}_{t|t}$ respectively. As such, they can be utilised to associate a measure of accuracy with the state estimates. From a computational perspective, it is worthwhile noting that $\boldsymbol{\Sigma}_{t|t-1}$ in eqn. (5.10b), the Kalman gain in eqn. (5.11) and $\boldsymbol{\Sigma}_{t|t}$ in eqn. (5.10d) are all independent of the observations, $\mathbf{y}_{1:t}$. Therefore, they can be computed in advance before the data sequence becomes available. Computation of $\boldsymbol{\Sigma}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t}$ prior to estimation of $\mathbf{x}_t$ allows for saving of computational time at estimation time and facilitates that the propagation of the MSE with time can be evaluated in advance.

As shown in Appendix B.2, the Kalman filter equations satisfy zero-mean estimation error and orthogonality to the measurements and hence constitute the optimal estimator of the states, $\mathbf{x}_t$, in the MSE sense. As such, the variance specified by the Cramér-Rao lower-bound (CRLB) can be achieved to obtain the best possible accuracy of the state estimates. Furthermore, the independence of the MSEof the observations and states (see Appendix B.2) allows for prior evaluation to estimation of the MSE in order to gauge in advance whether estimation of the states is worthwhile at all. Prior evaluation of the MSE is particularly helpful for where computational power is a scarce resource. A major benefit of the Kalman filter is the facilitation of sequential processing, allowing for real-time processing. The recursive structure of eqn. (5.10) is also computationally appealing as only the most recent estimates at time t are required in order to propagate the estimates to time t.

Although originally developed for linear, Gaussian processes, the Kalman filter

was shown to be optimal for several non-Gaussian systems as well. The linearity constraints can therefore be relieved by variants of the Kalman filter such as the extended Kalman filter [172–174] or the unscented Kalman filter [175–179]. Further Kalman filter variants of the Kalman filter exist, allowing for prediction of future states (using the Kalman predictor), smoothing of estimates by using past and future observations (using the fixed-interval smoother, the fixed-lag smoother, or the fixed-point smoother) or even estimation of continuous time systems (using the Kalman-Bucy filter).

However, the Kalman filter restricts the applicability to many signal processing problems in practice due to the assumption of known model parameters, i.e., known state transition and measurement transformation. In most signal processing applications – and in particular speech processing –, the states of the dynamic system need to be estimated blindly, i.e., without explicit prior knowledge of the model parameters. However, the performance of the Kalman filter degrades significantly for unknown model parameters, rendering the Kalman filter inappropriate for blind estimation problems.

Nonetheless, it is often desired to integrate Kalman filters as the optimal state estimator in blind estimation frameworks. Kalman filters can be integrated in possibly non-linear blind estimation frameworks by means of Rao-Blackwellisation. Whilst the model parameters are obtained using a separate estimator, their estimates are used for source signal estimation with the Kalman filter. Rao-Blackwellisation is explained in further detail sect. §5.6.

However, as will be shown in Chap. 6, expressions for certain system model parameters are not available in closed form such that, e.g., the MMSE estimate in eqn. (5.5) cannot be evaluated analytically. Therefore, the integral has to be *approximated* instead. Monte Carlo integration technique allow for the evaluation of the integral in eqn. (5.5) by drawing random samples from specified distributions. As will be discussed in Chap. 6, the analytically intractable parameters are often of large model order, with boundary conditions dependent on the order of the parameters, i.e., the intractable pdf must be evaluated for a large number of variables over a vast region of support. Monte Carlo techniques are particularly apt at solving multidimensional integrals with complicated boundary conditions and are therefore utilised in Chap. 6 for the evaluation of intractable parameters.

The remainder of this chapter reviews the necessary methodology required for the Monte Carlo techniques relevant to this thesis. Sect. §5.4.1 reviews the concept of perfect Monte Carlo integration, assuming that the posterior pdf is easy to sample

from, such that the integral can be solved by successively drawing samples from the posterior pdf. As the posterior pdf is often difficult or impossible to sample from, sect. §5.4.2 introduces acceptance-rejection sampling, drawing random variates from a proposal (or hypothesis) distribution and accepting the sampling if it lies within the proposal distribution, and rejecting it otherwise. Acceptance-rejection sampling is referred to when discussing the stability of parameters and how to ensure poles inside the unit circle. As acceptance-rejection sampling can lead to excessive over-sampling, sect. §5.4.3 introduces the concept of importance sampling, where samples are drawn from the hypothesis distribution and are weighted according to the discrepancy between the hypothesis distribution and posterior pdf and approximate the integral by means of a discrete sum over the weighted samples. Bayesian importance sampling, discussed in sect. §5.4.4, extends the posterior pdf in the importance sampling approach by means of Bayes's theorem into the likelihood function, prior pdf, and evidence term. Importance sampling and Bayesian importance sampling constitute the foundation for the SMC approaches discussed in sect. §5.5 and utilised in Chap. 6 for parameter estimation. As it is often desirable to *sequentially* evaluate integrals over the posterior pdf, sequential importance sampling techniques (sect. §5.5.1) update the posterior pdf at time t by means of its trajectory between 0 and $t-1$. The optimal choice of the hypothesis distribution the samples are drawn from is discussed in sect. §5.5.2. In order to retain statistically relevant samples only, resampling techniques as discussed in sect. §5.5.3 are utilised. Finally, Rao-Blackwellisation, where analyitcally tractable subspaces are estimated using their optimal estimators, whereas intractable parameters are obtained using sequential importance resampling is discussed in sect. §5.6. The discussion in this chapter is summarised in sect. §5.7.

5.4 Monte Carlo integration

Monte Carlo integration methods are sub-optimal methods that *numerically* solve the multidimensional integral

$$\boldsymbol{\varphi}_{0:t}^{\mathrm{MMSE}} = \int \boldsymbol{\varphi}_{0:t} p\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t} \tag{5.12}$$

by approximating the continuous integral by point-mass measures. In general, Monte Carlo approaches operate by 1. defining the region of support of the possible inputs, 2. drawing random samples from the region of support using predefined distributions, 3. combining the samples to generate the final estimate .

In their most naïve form, referred to as perfect Monte Carlo integration, Monte Carlo methods assume that it is possible to draw random samples from the posterior

pdf whose integral is sought.

5.4.1 Perfect Monte Carlo integration

Assuming that the posterior pdf is easy to sample from, N independent and identically distributed (i. i. d.) samples, $\boldsymbol{\varphi}_{0:t}^{(i)}$, $i \in \mathcal{N}$ of the desired variables, $\boldsymbol{\varphi}_{0:t}$, are drawn from the posterior pdf $p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})$. The posterior pdf can thus be approximated by the point-mass measure,

$$\tilde{p}(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}) = \frac{1}{N} \sum_{i=1}^{N} \delta\left(\boldsymbol{\varphi}_{0:t} - \boldsymbol{\varphi}_{0:t}^{(i)}\right). \tag{5.13}$$

The continuous expected value of $\boldsymbol{\varphi}_{0:t}$ in eqn. (5.5),

$$\hat{\boldsymbol{\varphi}}_{0:t} = \int \boldsymbol{\varphi}_{0:t} p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}) \, d\boldsymbol{\varphi}_{0:t}, \tag{5.5}$$

can thus be approximated by the discrete sum

$$\hat{\boldsymbol{\varphi}}_{0:t} = \int \boldsymbol{\varphi}_{0:t} \tilde{p}(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}) \, d\boldsymbol{\varphi}_{0:t} \approx \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\varphi}_{0:t}^{(i)} \tag{5.14}$$

According to the strong law of large numbers (SLLN), the average of an experiment performed for a large number of times is close to the expected value of the estimate [180], i.e.,

$$\lim_{N \mapsto \infty} \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\varphi}_{0:k}^{(i)} \mapsto \mathbb{E}_{p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})}[\boldsymbol{\varphi}_{0:t}] = \hat{\boldsymbol{\varphi}}_{0:t}$$

Furthermore, the variance satisfies, [25]:

$$\begin{aligned} \mathrm{var}_{\tilde{p}(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})}[\boldsymbol{\varphi}_{0:t}] &= \mathbb{E}_{\tilde{p}(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})}\left[(\boldsymbol{\varphi}_{0:t})^2\right] - \mathbb{E}^2_{\tilde{p}(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})}[\boldsymbol{\varphi}_{0:t}] \\ &= \int \|\boldsymbol{\varphi}_{0:t}\|^2 \tilde{p}(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}) \, d\boldsymbol{\varphi}_{0:t} - \|\hat{\boldsymbol{\varphi}}_{0:t}\|^2 \end{aligned}$$

which is *independent* of the state-space dimension as $N \to \infty$ as opposed to deterministic integration techniques whose convergence of the estimation error deteriorates with increasing dimensionality [25].

Perfect Monte Carlo integration assumes that the posterior pdf can be sampled from. However, for most distributions it is often difficult or even impossible to sample from the posterior density directly. More applicable Monte Carlo integration methods sample instead from a hypothesis distribution, $\pi(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})$, instead that is easy to sample from and approximates the posterior pdf. As the hypothesis distribution only

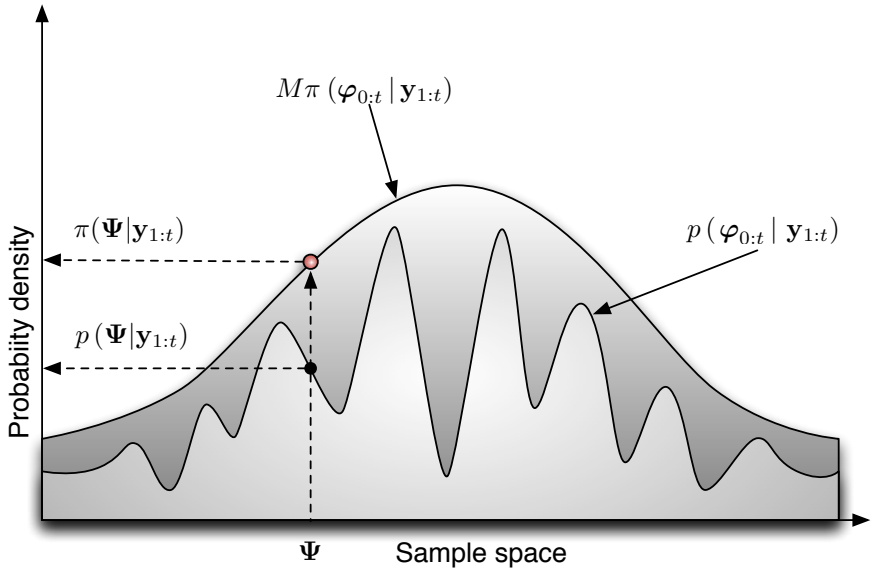


Figure 5.2: *Acceptance-Rejection sampling for posterior density,* $\mathrm{p}\,(\,\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})$*, and proposal density,* $\pi\,(\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})$*; Ratio of proposal and posterior density are compared for sample* .

approximates the posterior pdf, discrepancies between the hypothesis and proposal distribution exist. Random variates from $\pi\,(\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})$ thus do not always represent samples drawn from the posterior pdf. One approach to account for this discrepancy is to only accept samples that fall in the region of $\mathrm{p}\,(\,\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})$ and reject any other random variates. This approach is also known as acceptance-rejection sampling and is described in the following subsection.

5.4.2 Acceptance-Rejection sampling

Instead of sampling from the posterior pdf, acceptance-rejection sampling draws samples from a proposal distribution, $\pi\,(\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})$, that is easy to sample from and satisfies $\mathrm{M}\pi\,(\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}}) > \mathrm{p}\,(\,\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})$ for some scaling constant M (see Fig. 5.2) as illustrated in Fig. 5.2. As the hypothesis distribution represents a generous envelope distribution of the posterior pdf, one has to be wary of oversampling. On average, it can be expected that the too many of variates that take on a value $\boldsymbol{\Psi}$ are drawn by a factor of $\pi(\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})/\mathrm{p}(\,\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})$. To reduce the amount of oversampled variates, a number of variates proportional to the oversampling should be discarded. Therefore, samples will be accepted with probability

$$\Pr\,(\text{accept}\,\boldsymbol{\Psi}) = \frac{\mathrm{p}\,(\,\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})}{\mathrm{M}\pi\,(\boldsymbol{\varphi}_{0:\mathrm{t}}\,|\,\mathbf{y}_{1:\mathrm{t}})} \tag{5.15}$$

Initialization:
Select randomly or deterministically $\boldsymbol{\varphi}_{0:t}^{(1)}$;

Acceptance-Rejection iterations:
for $i = 2, \ldots, N$ **do**
  Generate a random sample from the proposal distribution:
  $\boldsymbol{\varphi}_{0:t}^{\star} \sim \pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$;
  Generate a random sample from the uniform distribution: $u \sim \mathcal{U}_{[0,\,1]}$;
  if $u \leq \frac{\pi(\boldsymbol{\varphi}_{0:t}^{\star} \,|\, \mathbf{y}_{1:t})}{Mp(\boldsymbol{\varphi}_{0:t}^{\star} \,|\, \mathbf{y}_{1:t})}$ **then**
    Accept proposed sample: $\boldsymbol{\varphi}_{0:t}^{(i)} = \boldsymbol{\varphi}_{0:t}^{\star}$;
  else
    Reject and go back to generating a random variate from the proposal distribution until sample accepted;
  end
end

Algorithm 5.1: Acceptance-Rejection sampling

and rejected otherwise. The estimate of the model parameters, $\hat{\boldsymbol{\theta}}_{0:t}$ is subsequently obtained by applying the Monte Carlo (MC) integration principle in eqn. (5.14) to the accepted candidates and scaling by the fraction of samples falling into the area of the desired posterior pdf.

However, it can be difficult even to find a proposal density that is easy to sample from such that the condition $M\pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t}) > p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$ holds. Furthermore, if the parameter-space is high-dimensional, the acceptance probability of the candidate is usually very small, leading to excessive over-sampling due to an undesirably large number of rejections.

5.4.3 Importance sampling

Also in importance sampling, the hypothesis distribution, also referred to as the importance function, $\pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$, is used to sample candidates, where $\pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$ is easy to sample from and approximates $p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$ with the same support [16, 24, 181, 182]. As opposed to acceptance-rejection sampling, *all* candidates are accepted. In order to account for discrepancies between the proposal distribution, $\pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$, and the desired posterior pdf, $p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$, samples are weighted by a function of the observations.

Assuming that the proposal distribution and the desired posterior have the same region of support, i.e., $\pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$ only takes values in the region support of $p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$, denoted as Φ, then $\int_{\Phi} \pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})\, d\boldsymbol{\varphi}_{0:t} = 1$. Thus, the proposal distribution can be

introduced to the posterior distribution by expanding to

$$\mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) = \frac{\mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)}{\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)} \pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) = w_t^{\star}\, \pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right), \tag{5.16}$$

where $w_t^{\star}$ are the so-called importance weights and determine the discrepancy between the posterior and the proposal distribution via

$$\mathbf{w}_{0:t}^{\star} \triangleq \frac{\mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)}{\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)}. \tag{5.17}$$

Using eqn. (5.16), the MMSE estimate of the unknowns, $\boldsymbol{\varphi}_{0:t}^{\mathrm{MMSE}}$, can be written as

$$\begin{aligned} \boldsymbol{\varphi}_{0:t}^{\mathrm{MMSE}} &= \int_{\Phi} \boldsymbol{\varphi}_{0:t} \mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t} = \int_{\Phi} \boldsymbol{\varphi}_{0:t} \frac{\mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)}{\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)} \pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t} \\ &= \mathbb{E}_{\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)}\left[\boldsymbol{\varphi}_{0:t}\, \mathbf{w}_{0:t}^{\star}\right]. \end{aligned} \tag{5.18}$$

if and only if $\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) \neq 0$ for any $\boldsymbol{\varphi}_{0:t} \in \Phi$ for which $\mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)$. Bearing in mind the principle of Perfect Monte Carlo simulation, assume N i. i. d. samples, $\boldsymbol{\varphi}_{0:t}^{(i)}$, $i \in \mathcal{N}$ can be drawn from the importance function. The continuous integral in eqn. (5.18) can thus be approximated by the discrete sum:

$$\boldsymbol{\varphi}_{0:t}^{\mathrm{MMSE}} \approx \widehat{\mathbf{f}}_{0:t}^{\mathrm{MMSE}} = \sum_{i \in \mathcal{N}} \frac{\boldsymbol{\varphi}_{0:t}^{(i)}\, \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}^{(i)} \,\middle|\, \mathbf{y}_{1:t}\right)}{\pi\left(\boldsymbol{\varphi}_{0:t}^{(i)} \mid \mathbf{y}_{1:t}\right)} = \sum_{i \in \mathcal{N}} \boldsymbol{\varphi}_{0:t}^{(i)}\, w_{0:t}^{\star\,(i)}. \tag{5.19}$$

In order to perform importance sampling, a closed form expression of the posterior pdf must be available in order to solve the importance weights in eqn. (5.17). However, recalling eqns. (5.1) and (5.2) on page 79 the parameters must be linear in the observations in order to solve the integral required for the evidence term,

$$\mathrm{p}\left(\mathbf{y}_{1:t}\right) = \int_{\Phi} \mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t}. \tag{5.2}$$

In most applications the solution of eqn. (5.2) is not possible due to non-linear dependencies of the parameters in the observations. To circumvent evaluation of the evidence term, Bayesian importance sampling exploits the fact that $\mathrm{p}\left(\mathbf{y}_{1:t}\right)$ is independent of $\boldsymbol{\varphi}_{0:t}$, thus acting as a scaling constant only.

5.4.4 Bayesian importance sampling

Where the evidence term, $\mathrm{p}\left(\mathbf{y}_{1:t}\right)$, is intractable, Bayesian importance provides a reformulation of the importance sampling framework resulting in the cancellation of the evidence altogether. By application of Bayes's theorem, the posterior pdf can be

rewritten as

$$\mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) = \frac{\mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right)}{\mathrm{p}\left(\mathbf{y}_{1:t}\right)} \tag{5.1}$$

Inserting this into eqn. (5.17), the importance weights are given by

$$w^{\star}_{0:t} \triangleq \frac{\mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)}{\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)} = \frac{\mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right)}{\mathrm{p}\left(\mathbf{y}_{1:t}\right) \pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)}. \tag{5.20}$$

As the evidence term, $\mathrm{p}\left(\mathbf{y}_{1:t}\right)$, is independent of the unknown variable, $\boldsymbol{\varphi}_{0:t}$, it thus acts as a scaling constant to the weights. As $\mathrm{p}\left(\mathbf{y}_{1:t}\right)$ is assumed intractable, the importance weights are redefined by neglecting the normalising evidence, thus resulting in the *unnormalised* importance weights, $w_{0:t}$, where

$$w_{0:t} \triangleq \frac{\mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right)}{\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)}. \tag{5.21}$$

The MMSE estimate of $\boldsymbol{\varphi}_{0:t}$ can thus be derived slightly differently from eqn. (5.18) by inserting eqn. (5.1) into eqn. (5.18), i.e.,

$$\begin{aligned}
\boldsymbol{\varphi}^{\mathrm{MMSE}}_{0:t} &= \int_{\Phi} \boldsymbol{\varphi}_{0:t}\, \mathrm{p}\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t} = \int_{\Phi} \boldsymbol{\varphi}_{0:t}\, \frac{\mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right)}{\mathrm{p}\left(\mathbf{y}_{1:t}\right)} \mathrm{d}\boldsymbol{\varphi}_{0:t} \\
&= \frac{\int_{\Phi} \boldsymbol{\varphi}_{0:t}\, \mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t}}{\mathrm{p}\left(\mathbf{y}_{1:t}\right)} = \frac{\int_{\Phi} \boldsymbol{\varphi}_{0:t}\, \mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t}}{\int_{\Phi} \mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t}}
\end{aligned}$$

which, using $\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)$, can be extended to

$$= \frac{\int_{\Phi} \boldsymbol{\varphi}_{0:t}\, \frac{\mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right)}{\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)} \pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t}}{\int_{\Phi} \frac{\mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right)}{\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)} \pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t}}. \tag{5.22}$$

and by inserting eqn. (5.21) can be simplified to

$$= \frac{\int_{\Phi} \boldsymbol{\varphi}_{0:t}\, w_{0:t}\, \pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t}}{\int_{\Phi} w_{0:t}\, \pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right) \mathrm{d}\boldsymbol{\varphi}_{0:t}} = \frac{\mathbb{E}_{\pi(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})}\left[\boldsymbol{\varphi}_{0:t}\, w_{0:t}\right]}{\mathbb{E}_{\pi(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})}\left[w_{0:t}\right]}. \tag{5.23}$$

Due to the independence of $\boldsymbol{\varphi}_{0:t}$, the evidence term, $\mathrm{p}\left(\mathbf{y}_{1:t}\right)$, thus cancelled from the MMSE estimate, such that only the tractable likelihood function, $\mathrm{p}\left(\mathbf{y}_{1:t} \mid \boldsymbol{\varphi}_{0:t}\right)$, the prior pdf, $\mathrm{p}\left(\boldsymbol{\varphi}_{0:t}\right)$, and the hypothesis distribution, $\pi\left(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}\right)$, are involved in com-

puting $\boldsymbol{\varphi}_{0:t}^{\mathrm{MMSE}}$. Similar to eqn. (5.19), if N i. i. d. samples, $\boldsymbol{\varphi}_{0:t}^{(i)}$, $i \in \mathcal{N}$ are drawn from the importance function, $\boldsymbol{\varphi}_{0:t}^{\mathrm{MMSE}}$ can be approximated by

$$\boldsymbol{\varphi}_{0:t}^{\mathrm{MMSE}} \approx \hat{\boldsymbol{\varphi}}_{0:t} = \frac{\frac{1}{N}\sum_{i\in\mathcal{N}} \boldsymbol{\varphi}_{0:t}^{(i)} w_{0:t}^{(i)}}{\frac{1}{N}\sum_{j\in\mathcal{N}} w_{0:t}^{(j)}} = \sum_{i\in\mathcal{N}} \boldsymbol{\varphi}_{0:t}^{(i)} \tilde{w}_{0:t}^{(i)} \tag{5.24}$$

where the particles are re-normalised via

$$\tilde{w}_{0:t}^{(i)} \triangleq \frac{w_{0:t}^{(i)}}{\sum_{j\in\mathcal{N}} w_{0:t}^{(j)}}. \tag{5.25}$$

Importance sampling as described above requires all data, $\mathbf{y}_{1:t}$, before estimating the posterior pdf, $p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})$, and hence the MMSE estimate, $\hat{\boldsymbol{\varphi}}_{0:t}$. Thus, as a new sample, $\mathbf{y}_{t+1}$, becomes available, the importance weights in eqn. (5.21) need to be recomputed over the entire time trajectory. As the trajectory of $\mathbf{y}_{1:t}$ and $\boldsymbol{\varphi}_{0:t}$ increases in dimension with time, the computational complexity involved in computing $w_{0:t}$ from eqn. (5.21) increases with time.

Instead of performing importance sampling on a batch of data, $p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})$ can be estimated recursively without modifying the trajectory of previously simulated samples, $\boldsymbol{\varphi}_{0:t-1}$ using *sequential* importance sampling.

5.5 SMC methods and particle filters

In practice, it is often desirable to take advantage of the fact that dynamic models change with time and can therefore be *updated* with time. The posterior pdf can thus be rewritten *sequentially* using Bayes's theorem as

$$\begin{aligned} p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t}) &= \frac{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t})\, p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t-1})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} \\ &= \frac{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t})\, p(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1})\, p(\boldsymbol{\varphi}_{1:t-1} \mid \mathbf{y}_{1:t-1})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})} \\ &= p(\boldsymbol{\varphi}_{0:t-1} \mid \mathbf{y}_{1:t-1}) \times \frac{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t})\, p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})}. \end{aligned} \tag{5.26}$$

The posterior pdf $p(\boldsymbol{\varphi}_{0:t} \mid \mathbf{y}_{1:t})$ thus takes the previous posterior, $p(\boldsymbol{\varphi}_{0:t-1} \mid \mathbf{y}_{1:t-1})$, and updates it using knowledge acquired from the most immediate observations, $\mathbf{y}_t$. Based on eqn. (5.26), importance sampling can reformulated in a sequential framework, also referred to as sequential importance sampling (SIS).

SIS forms the basis of a subset of SMC methods that was developed since the 1950s

and is known under various names such as bootstrap filtering [23], conditional density propagation (CONDENSATION) [183], survival of the fittest [184], interacting particle approximations [185], or, most commonly, particle filtering [16]. Although particle filters are often used synonymously with SMC, SMC methods include any estimator that propagates the point-mass measure of the random variates drawn from the hypothesis distribution through time [171].

Particle filters perform SMC estimation by means of SIS. The key idea is thus to represent the desired posterior pdf by a cloud of random variates (also referred to as particles) drawn from the hypothesis distribution and associated with weights. Estimates are computed based on the particles and their weights. As the number of particles increases, the point mass distribution becomes an accurate estimate of the posterior pdf and the SIS filter approximates the optimal Bayesian estimator. An extensive review of SMC methods is edited by Doucet *et al.* [16]. Introductory video lectures on the topic can be found online, e.g., given by Doucet[1] or de Freitas[2].

5.5.1 Sequential importance sampling

In order to facilitate a sequential formulation of the importance sampling framework, the proposal distribution is assumed to be i. i. d., i.e.,

$$\pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t}) = \pi(\boldsymbol{\varphi}_0) \prod_{k=1}^{t} \pi(\boldsymbol{\varphi}_k \,|\, \boldsymbol{\varphi}_{0:k-1}, \mathbf{y}_{1:k}) \tag{5.27}$$

$$= \pi(\boldsymbol{\varphi}_{0:t-1} \,|\, \mathbf{y}_{1:t-1})\, \pi(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})\,. \tag{5.28}$$

In other words, the importance function, $\pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$, admits the importance function at $t-1$, $p(\boldsymbol{\varphi}_{0:t-1} \,|\, \mathbf{y}_{1:t-1})$. In this case, the importance weights in eqn. (5.21) can be re-formulated using eqns. (5.26) and (5.27) as

$$\begin{aligned} w_{0:t} &= \frac{p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})}{\pi(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})}\, p(\boldsymbol{\varphi}_{0:t}) \\ &= \frac{p(\boldsymbol{\varphi}_{0:t-1} \,|\, \mathbf{y}_{1:t-1})}{\pi(\boldsymbol{\varphi}_{0:t-1} \,|\, \mathbf{y}_{1:t-1})} \times \frac{p(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t})\, p(\boldsymbol{\varphi}_t \,|\, \boldsymbol{\varphi}_{0:t-1})}{p(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1})\, \pi(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})} p(\boldsymbol{\varphi}_t \,|\, \boldsymbol{\varphi}_{0:t-1})\, p(\boldsymbol{\varphi}_{0:t-1}) \\ &= w_{0:t-1} \times \underbrace{\frac{p(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t})\, p(\boldsymbol{\varphi}_t \,|\, \boldsymbol{\varphi}_{0:t-1})}{\pi(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})}}_{w_t} \end{aligned} \tag{5.29}$$

[1] `http://videolectures.net/mlss07_doucet_smcm/`
[2] `http://videolectures.net/mlss08au_freitas_asm/`

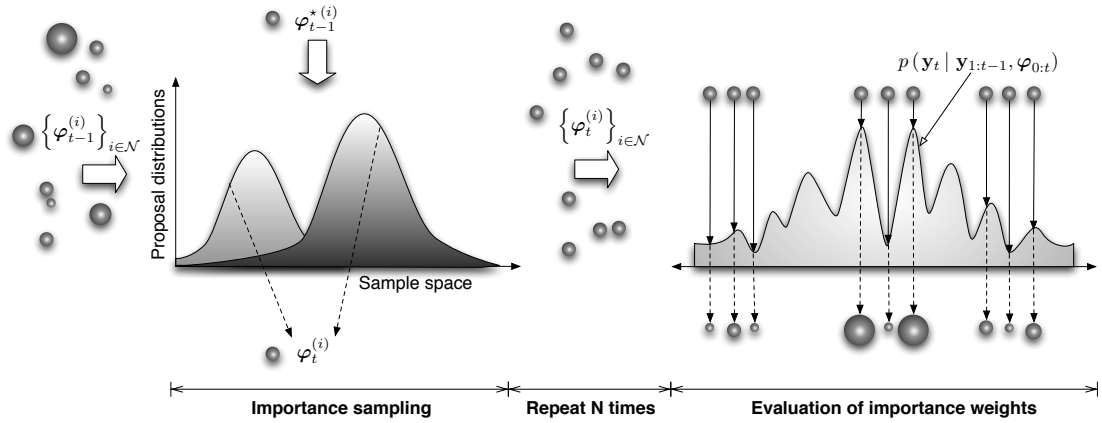


Figure 5.3: *Sequential importance sampling, where particles are drawn from the importance distribution and weights are evaluated according to the likelihood function.*

Recalling that the posterior density can be approximated by the discrete weighted approximation:

$$\begin{aligned} \mathrm{p}\left(\boldsymbol{\varphi}_{0:\mathrm{t}} \mid \mathbf{y}_{1:\mathrm{t}}\right) &= \sum_{\mathrm{i}\in\mathcal{N}} \tilde{w}_{0:\mathrm{t}}^{(\mathrm{i})}\, \delta\left(\boldsymbol{\varphi}_{0:\mathrm{t}} - \boldsymbol{\varphi}_{0:\mathrm{t}}^{(\mathrm{i})}\right) \\ &= \sum_{\mathrm{i}\in\mathcal{N}} \tilde{w}_{0:\mathrm{t}-1}^{(\mathrm{i})}\, \tilde{w}_{\mathrm{t}}^{(\mathrm{i})}\, \delta\left(\begin{bmatrix}\boldsymbol{\varphi}_{\mathrm{t}}\\ \vdots \\ \boldsymbol{\varphi}_{0}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\varphi}_{\mathrm{t}}^{(\mathrm{i})}\\ \vdots \\ \boldsymbol{\varphi}_{0}^{(\mathrm{i})}\end{bmatrix}\right) \\ &= \sum_{\mathrm{i}\in\mathcal{N}} \tilde{w}_{0:\mathrm{t}}^{(\mathrm{i})} \delta\left(\boldsymbol{\varphi}_{0:\mathrm{t}-1} - \boldsymbol{\varphi}_{0:\mathrm{t}-1}^{(\mathrm{i})}\right) \sum_{\mathrm{j}\in\mathcal{N}} \tilde{w}_{\mathrm{t}}^{(\mathrm{j})}\, \delta\left(\boldsymbol{\varphi}_{\mathrm{t}} - \boldsymbol{\varphi}_{\mathrm{t}}^{(\mathrm{j})}\right). \end{aligned} \tag{5.30}$$

where the posterior pdf for the variables at time t is

$$\mathrm{p}\left(\boldsymbol{\varphi}_{\mathrm{t}} \mid \mathbf{y}_{1:\mathrm{t}}, \boldsymbol{\varphi}_{0:\mathrm{t}-1}\right) = \sum_{\mathrm{i}\in\mathcal{N}} \tilde{w}_{\mathrm{t}}^{(\mathrm{i})}\, \delta(\boldsymbol{\varphi}_{\mathrm{t}} - \boldsymbol{\varphi}_{\mathrm{t}}^{(\mathrm{i})}). \tag{5.31}$$

Likewise, the trajectory of estimates is obtained via eqn. (5.24) on page 93 as

$$\hat{\boldsymbol{\varphi}}_{0:\mathrm{t}} = \sum_{\mathrm{i}\in\mathcal{N}} \tilde{w}_{0:\mathrm{t}}^{(\mathrm{i})}\, \boldsymbol{\varphi}_{0:\mathrm{t}}^{(\mathrm{i})} = \sum_{\mathrm{i}\in\mathcal{N}} \tilde{w}_{0:\mathrm{t}-1}^{(\mathrm{i})} \tilde{w}_{\mathrm{t}}^{(\mathrm{i})} \begin{bmatrix}\boldsymbol{\varphi}_{\mathrm{t}}^{(\mathrm{i})}\\ \boldsymbol{\varphi}_{0:\mathrm{t}-1}^{(\mathrm{i})}\end{bmatrix}. \tag{5.32}$$

Thus, rather than estimating $\hat{\boldsymbol{\theta}}_{0:\mathrm{t}}$ in one batch, the current estimate can be appended to the trajectory at the previous time step via $\hat{\boldsymbol{\varphi}}_{0:\mathrm{t}} = \{\hat{\boldsymbol{\varphi}}_{0:\mathrm{t}-1},\ \hat{\boldsymbol{\varphi}}_{\mathrm{t}}\}$, where

$$\hat{\boldsymbol{\varphi}}_{\mathrm{t}} = \sum_{\mathrm{i}\in\mathcal{N}} \tilde{w}_{\mathrm{t}}^{(\mathrm{i})}\, \boldsymbol{\varphi}_{\mathrm{t}}^{(\mathrm{i})}. \tag{5.33}$$

SIS is summarised in Alg. 5.2 and illustrated in Fig. 5.3.

for $t = 1, 2, 3, \ldots$ **do**
  for $i = 1, \ldots, N$ **do**
    Draw the importance samples $\boldsymbol{\varphi}_t^{(i)} \sim \pi\left(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{t-1}^{(i)}\right)$;
    Evaluate the weights up to a scaling constant:

$$w_{0:t} = w_{0:t-1} \times \frac{p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t}\right) p\left(\boldsymbol{\varphi}_t \,|\, \boldsymbol{\varphi}_{0:t-1}\right)}{\pi\left(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right)} \tag{5.29}$$

  end
  for $i = 1, \ldots, N$ **do**
    Normalise the weights (eqn. (5.25)):

$$\tilde{w}_{0:t}^{(i)} \triangleq w_{0:t}^{(i)} \Big/ \sum_{j \in \mathcal{N}} w_{0:t}^{(j)}. \tag{5.25}$$

  end
end

Algorithm 5.2: sequential importance sampling

5.5.2 Choice of importance sampling function

The performance of SIS approaches is highly dependent on the choice of importance function. The optimal importance function was first derived by Zaritskii *et al.* in [186], extended to a special case by Akashi and Kumamoto in [187] and utilised in [188–190]. The optimal importance function minimises the variance upon the samples $\boldsymbol{\varphi}_{0:t}^{(i)}$, $i \in \mathcal{N}$ and the observations, $\mathbf{y}_{1:t}$ [24]. As shown in Appendix B.4, the variance of the estimator can be expressed as

$$\begin{aligned} &\mathrm{var}_{\pi(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})}\left[w_t\right] \\ &= w_{t-1}^2 \left[\int \frac{\left(p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t}\right) p\left(\boldsymbol{\varphi}_t \,|\, \boldsymbol{\varphi}_{0:t-1}\right)\right)^2}{\pi\left(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right)} \mathrm{d}\boldsymbol{\varphi}_t - p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}\right)^2\right]. \end{aligned} \tag{5.34}$$

In order to obtain zero variance, the proposal distribution is of the form (see Appendix B.4 or [191])

$$\pi\left(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right) = p\left(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right). \tag{5.35}$$

The optimal importance weights are found by inserting eqn. (5.35) into eqn. (5.29), i.e.,

$$\begin{aligned} w_{0:t} &= w_{0:t-1} \times \frac{p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t}\right) p\left(\boldsymbol{\varphi}_t \,|\, \boldsymbol{\varphi}_{0:t-1}\right)}{\pi\left(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right)} \qquad (5.29) \\ &= w_{:t-1} \times \frac{p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t}\right) p\left(\boldsymbol{\varphi}_t \,|\, \boldsymbol{\varphi}_{0:t-1}\right)}{p\left(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right)} \end{aligned}$$

and, by application of Bayes's rule in the denominator,

$$\begin{aligned} w_{0:t} &= w_{0:t-1} \times \frac{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t})\, p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t})\, p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1})} \times p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}) \\ &= w_{0:t-1} \times p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}) \end{aligned} \tag{5.36}$$

The likelihood term, $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1})$, can theoretically be evaluated via,

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}) = \int p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t})\, p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1})\, d\boldsymbol{\varphi}_t. \tag{5.37}$$

However, as discussed in sect. §5.4.3, in general, the integral is analytically intractable in practice due to a non-linear relation between $\boldsymbol{\varphi}_t$ and the likelihood. In those cases, the optimal importance weights in eqn. (5.36) cannot be evaluated analytically and the importance function cannot be sampled from. Approximation of the optimal importance function is thus necessary instead. Prior importance sampling, the most common form of sub-optimal importance sampling, is discussed in the following.

5.5.2.1 Prior importance sampling

In cases where the optimal importance function cannot be evaluated, the prior pdf of the unknown variables, $p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1})$, can be used as the importance function. The approach was initially proposed by Handschin [192] and Handschin and Mayne [193] and more recently applied by, e.g., Tanizaki and Mariano in [20, 194]. Inserting the prior pdf into the importance weights in eqn. (5.29), i.e., $\pi(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}) = p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1})$, the weights for the prior importance sampling scheme are given by

$$w_t = w_{t-1} \times \frac{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t})\, p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1})}{\pi(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})} \tag{5.29}$$

$$= w_{t-1} \times p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t}) \tag{5.38}$$

In essence, the importance weights evaluate the quality of the estimates. The samples of $\boldsymbol{\varphi}_t$ are compared to the underlying known truth in the observations in order evaluate how realistic their choices are. Likely samples subsequently get assigned high weights, whilst unlikely samples are low-weighted.

5.5.2.2 Issues with non-optimal importance functions

A main problem with non-optimal importance sampling, however, is that the hypothesis distribution only approximates $p(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})$, thus leading to a discrepancy between the samples and the actual values of the hidden states. Due to the discrepancy between the hypothesis and posterior pdf, the variance of the importance weights increases stochastically with time as shown by Kong in [189]. Thus, the discrepancy be-

tween samples drawn from the proposal distribution and actual values of the posterior pdf increases with time. As the random variates drawn from the hypothesis distribution become increasingly inaccurate, their weights decrease. After a few iterations, all but one importance weight are close to zero. Computational effort is thus dissipated to tracking sample trajectories whose weight prohibits them from contributing to the final estimate. This effect is known as the *degeneracy* of SIS. The effects of degeneracy by replacing particles with low weights with particles associated with high weights using so-called resampling schemes.

5.5.3 Resampling for avoidance of particle degeneracy

Resampling schemes [3] (also known as rejuvenation [190]) ensure that only statistically relevant samples are retained by replacing samples with low weights with samples associated with high weights. In other words, the posterior pdf at time t in eqn. (5.31),

$$p(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}) = \sum_{i \in \mathcal{N}} \tilde{w}_t^{(i)} \delta(\boldsymbol{\varphi}_t - \boldsymbol{\varphi}_t^{(i)}). \tag{5.31}$$

is replaced by [195]

$$p^\star(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}) = \sum_{i \in \mathcal{N}} \frac{1}{N} \delta(\boldsymbol{\varphi}_t - \boldsymbol{\varphi}_t^{\star(i)}). \tag{5.39}$$

where $\left\{ \boldsymbol{\varphi}_t^{\star(i)} \right\}_{i \in \mathcal{N}}$ are the resampled particles. As the weights are reset to equiprobable weights after each resampling step, eqn. (5.29) on page 94 reduce to

$$w_{0:t} = w_{0:t-1} \frac{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t}) \, p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1})}{\pi(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})} \tag{5.29}$$

$$\propto \frac{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t}) \, p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1})}{\pi(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})} = w_t \tag{5.40}$$

Resampling schemes therefore reselect the particle positions and weights such that the discrepancy between the resampled weights is reduced [196]. Particles can be resampled using stochastic resampling schemes or deterministic resampling. Following the argument in [197], stochastic sampling is analogous to spinning a roulette wheel where each particle corresponds to a pocket on the wheel. By repeatedly spinning the wheel, N particles chosen. Deterministic resampling schemes, in contrast, order the particles by their associated weight and choose the particles with biggest weights to replace those with small weights.

The four most frequently encountered and also most basic resampling schemes

Initialise the cdf: $c_1 = 0$;
for $i = 2, \ldots, N$ **do**
  Construct the cdf: $c_i = c_{i-1} + w_t^{(i)}$;
end
Draw a starting point: $u_1 \sim \mathcal{U}_{[0,\,1]}$;
for $j = 1, \ldots, N$ **do**
  Move along the cdf: $u_j = \frac{u_1 + j - 1}{N}$;
  while $u_j > c_i$ **do**
    $i = i + 1$;
  end
  Assign sample: $\boldsymbol{\varphi}_{0:t}^{(j)\star} = \boldsymbol{\varphi}_{0:t}^{(i)}$.
end

Algorithm 5.3: Systematic resampling according to [2].

are multinomial, systematic, and residual resampling. All four aim to enforce i. i. d. samples of the particles from the point-mass distribution in eqn. (5.31) by a process similar to the inverse transform [198, Chap. 2.1.2]. Random samples of the particle indices are drawn from a uniform distribution and transformed to a desired and known cummulative distribution function (cdf) via

$$X = F^{-1}(Y) \tag{5.41}$$

where $Y \sim \mathcal{U}_{[0,\,1]}$ and X is a continuous random variable with cdf $F(\cdot)$. Essentially, the cdf generated by the cumulative sum of the weighted particles is compared to a function of uniform variables.

5.5.3.1 Multinomial resampling

In multinomial resampling, N ordered uniform random numbers from a set of uniform samples, $\tilde{u}_j \sim \mathcal{U}_{[0,\,1]}$, i.e.,

$$u_j = u_{j+1}\,\tilde{u}_j^{1/j} \qquad\qquad u_N = \tilde{u}_N^{1/N},$$

N i. i. d. particles with uniform weights can be drawn from a multinomial distribution:

$$\boldsymbol{\varphi}_t^{\star\,(j)} = \boldsymbol{\varphi}_t^{(F^{-1}(u_j))}. \tag{5.42}$$

As multinomial resampling performs the inverse transform to N independent random samples, it can become computationally expensive. Systematic, stratified, and residual resampling improve the efficiency of the algorithm by dependent samples from a uniform distribution.

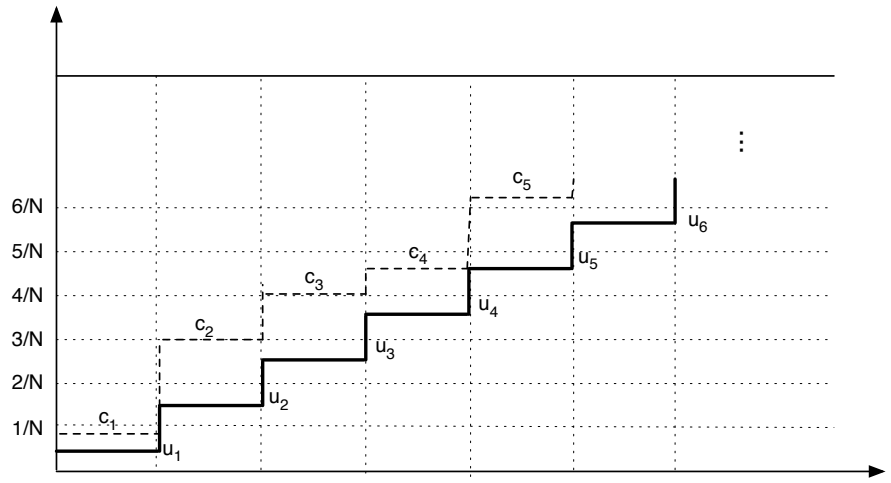


Figure 5.4: *Systematic resampling with resulting indices i = 2, 3, 4, 4, 6.*

5.5.3.2 Systematic resampling

Systematic resampling partitions the interval $(0, 1]$ into N disjoint and contiguous sets, i.e., it constructs an N-step ladder between 0 and 1 of equal step sizes, i.e.,

$$u_j = \frac{u_1 + N - 1}{N} \qquad u_1 \sim \mathcal{U}_{[0,\,1]}. \tag{5.43}$$

The particles are redrawn from the multinomial distribution in eqn. (5.42). Therefore, the N-step cdf of the weights is constructed and in each segment compared to the uniform ladder. If the ladder is greater than the cdf value the index is increased until the cdf is larger, otherwise the sample is assigned with the current index. Systematic resampling is summarised in Alg. 5.3 and illustrated in Fig. 5.4.

5.5.3.3 Residual resampling

For residual resampling, as used in [199], the number of replications of a particle is determined by rounding the product of the number of particles and the weights [3]. In a first step of the algorithm, the number of replications is computed and the weights normalised. Due to the calculations involved, the first step cannot guarantee that the number of particles is N. Thus, if any residual particles remain, the remaining resampling indices are drawn using systematic resampling in the second step and the weights are adjusted accordingly. Residual resampling is summarised in Alg. 5.4.

As discussed in [195], in addition to its ease of implementation, systematic resampling generates the lowest discrepancy between the weights of the particles and is requires the least computational expense as compared to multinomial and residual resampling. Systematic resampling is therefore widely used in the literature (see,

Initialise the residual number of particles: $N_r = N$;
for $i = 1, \ldots, N$ **do**
  Compute the index: $j(i) = \lfloor w_t^{(i)} \cdot N \rfloor$;
  Reassign weights: $w_t^{(i)} = w_t^{(i)} \cdot N - j(i)$;
  Update residual number of particles: $N_r = N - j(i)$;
end
if $M_r > 0$ **then**
  for $i = 1, \ldots, N$ **do**
    Renormalise weights: $w_t^{(i)} = w_t^{(i)} / N_r$;
  end
  Draw the M_r residual indices using systematic resampling: $j_r = \mathrm{SR}(M_r)$;
  for $i = 1, \ldots, N$ **do**
    Update the resampling indices: $j = j + j_r$;
    Assign sample: $\boldsymbol{\varphi}_{0:t}^{(i)\star} = \boldsymbol{\varphi}_{0:t}(j)$.
  end
else
  for $i = 1, \ldots, N$ **do**
    Assign sample: $\boldsymbol{\varphi}_{0:t}^{(i)\star} = \boldsymbol{\varphi}_{0:t}(j(i))$.
  end
end

Algorithm 5.4: Residual resampling according to [3].

e.g., [2]) and is used in this thesis as the preferred resampling scheme.

Although resampling avoids degeneracy of particles, it introduces several other issues: i) The parallelisability of particle filtering is reduced as the particles need to be combined for computation of the effective sample size and resampling indices; ii) As particles with high weight are statistically selected several times, diversity amongst particles is reduced and the particle cloud contains multiple copies of the same particles. For small process noise, all particles collapse to a single sample within few iterations. This effect is also known as sample impoverishment [2, 17]; iii) Since the diversity of particle paths is reduced, smoothed estimates of the particles' paths degenerate. Several approaches for avoidance of sample impoverishment have been proposed in the literature, including backward-forward filtering, Markov chain Monte Carlo (MCMC) move steps or resample-move algorithms [132, 200, 201]. As an additional MCMC move step in the particle filter framework can lead to significant increase in computational complexity, it is computationally more feasible to evaluate a measure of degeneracy of the particles at each time step. Only if the measure lies below a certain threshold, i.e., if the particles are degenerate, resampling is performed.

for $t = 1, 2, 3, \ldots$ **do**
  for $i = 1, \ldots, N$ **do**
    Draw the importance samples $\boldsymbol{\varphi}_t^{(i)} \sim \pi\left(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{t-1}^{(i)}\right)$;
    Evaluate the weights up to a scaling constant (eqn. (5.29));
  end
  for $i = 1, \ldots, N$ **do**
    Normalise the weights (eqn. (5.25)):
  end
  Compute the effective sample size (eqn. (5.44));
  if $\widehat{N}_{eff} < \textit{threshold}$ **then**
    Resample using, e.g., Alg. 5.3 or Alg. 5.4;
  end
end

Algorithm 5.5: sequential importance resampling

5.5.3.4 Degeneracy measure and deciding when to resample

Although resampling can be performed at any stage of the importance sampling scheme, it often adds computational burden and diversity amongst samples can be lost. On the other hand, if resampling is performed too rarely, the Monte Carlo scheme will be inefficient due to sample degeneracy. Resampling schemes thus evaluate a measure of degeneracy that determines whether resampling is necessary at that particular stage. Kong *et al*. [189] and Liu [202] propose to use the effective sample size, N_{eff}, as a measure of degeneracy, i.e.,

$$N_{eff} = \frac{N}{1 + \mathrm{var}_{\pi(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})}\left[w_t^\star\right]} = \frac{N}{\mathbb{E}_{\pi(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})}\left[\left(w_t^\star\right)^2\right]} \leq N.$$

A formal proof can be found in, e.g., [189]. Although N_{eff} cannot be evaluated exactly, an estimate, $\widehat{N}_{eff}$, can be obtained by

$$\widehat{N}_{eff} = \frac{1}{\sum_{i \in \mathcal{N}} \left(\tilde{w}_t^{(i)}\right)^2}. \tag{5.44}$$

If $\widehat{N}_{eff}$ is below a certain threshold, the particles are resampled. Incorporation of the resampling step in the SIS framework in Alg. 5.2 on page 96 is outlined in Alg. 5.5 and illustrated in Fig. 5.3. The resulting algorithm is known as sequential importance resampling (SIR).

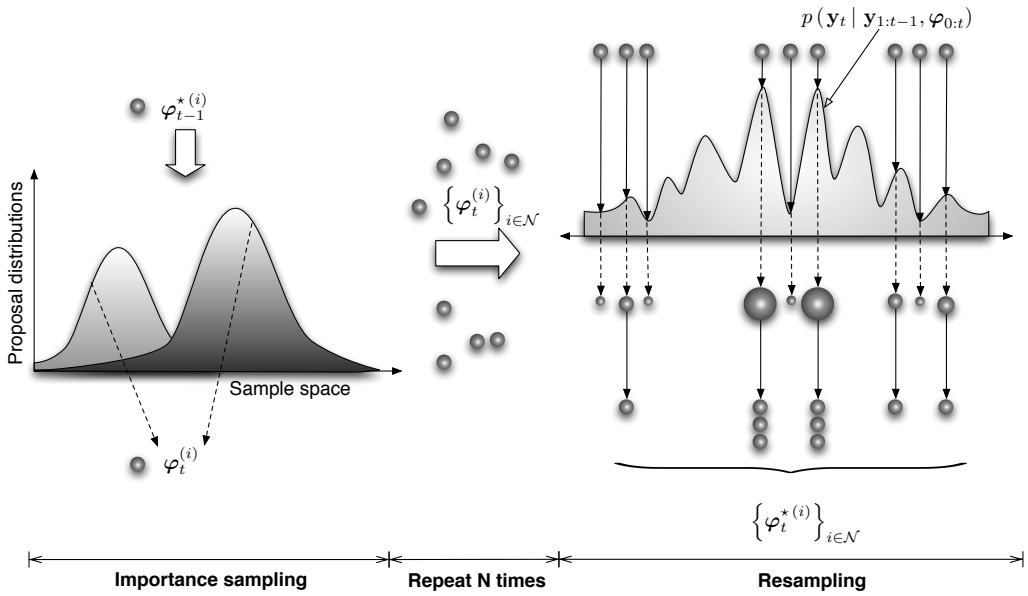


Figure 5.5: *Sequential importance resampling, where particles are drawn from the importance distribution and weights are determined by the likelihood function. Particles with low weights are replaced by particles associated with high weights.*

5.6 Rao-Blackwellisation of particle filters

In many estimation problems, the unknown variable, $\boldsymbol{\varphi}_{0:t}$, is the collection of several unknown parameters. Recalling the CGSS in eqn. (5.6) on page 83, and assuming that the model parameters are, in fact, unknown, the set of all unknown variables, $\boldsymbol{\varphi}_{0:t}$, is the *augmented* vector $\boldsymbol{\varphi}_{0:t} = \left[\boldsymbol{\theta}_{0:t}^{\mathsf{T}}, \mathbf{x}_{0:t}^{\mathsf{T}}\right]^{\mathsf{T}}$. Recalling that the states, $\mathbf{x}_{0:t}$, are analytically tractable and its optimal estimate can be obtained using the Kalman filter, it is desirable to estimate $\boldsymbol{\theta}_{0:t}$ and $\mathbf{x}_{0:t}$ separately.

Rao-Blackwellisation facilitates the separate estimation of variables in the same state space by marginalisation of analytical substructures as introduced by Casella in Roberts in [26]. The basic idea is based on the the Rao-Blackwell theorem, which relates the variance between a conditional and unconditional density, i.e.,

$$\operatorname{var}\left[\boldsymbol{\varphi}_{0:t}^{\text{MMSE}}\right] = \operatorname{var}\left[\mathbb{E}\left\{\boldsymbol{\theta}_{0:t}^{\text{MMSE}}, \mathbf{x}_{0:t}^{\text{MMSE}} \,\middle|\, \mathbf{x}_{0:t}\right\}\right] + \mathbb{E}\left[\operatorname{var}\left[\boldsymbol{\theta}_{0:t}^{\text{MMSE}}, \mathbf{x}_{0:t}^{\text{MMSE}} \,\middle|\, \mathbf{x}_{0:t}\right]\right]$$

Thus, by marginalising the analytically tractable solution of $\mathbf{x}_{0:t}$ from the non-tractable $\boldsymbol{\theta}_{0:t}$, the variance of the estimate, $\mathbb{E}\left\{\boldsymbol{\theta}_{0:t}^{\text{MMSE}}, \mathbf{x}_{0:t}^{\text{MMSE}} \,\middle|\, \mathbf{x}_{0:t}\right\}$, is improved by a non-negative term, $\mathbb{E}\left[\operatorname{var}\left[\boldsymbol{\theta}_{0:t}^{\text{MMSE}}, \mathbf{x}_{0:t}^{\text{MMSE}} \,\middle|\, \mathbf{x}_{0:t}\right]\right]$, as compared to the combined estimator $\boldsymbol{\varphi}_{0:t}^{\text{MMSE}}$. The variance of combined parameter estimation using particle filters can thus be reduced by marginalising analytically tractable substructures from the joint param-

eter space.

To show this, consider that the posterior pdf of all unknowns can be written by application of the probability chain rule as

$$\begin{aligned} \mathrm{p}\left(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right) &= \mathrm{p}\left(\boldsymbol{\theta}_t, \mathbf{x}_t \mid \mathbf{y}_{1:t}, \mathbf{x}_{0:t-1}, \boldsymbol{\theta}_{0:t-1}\right) \\ &= \mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \mathbf{x}_{0:t-1}, \boldsymbol{\theta}_{0:t}\right) \mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \mathbf{x}_{0:t-1}, \boldsymbol{\theta}_{0:t-1}\right) \\ &= \mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t}, \mathbf{x}_{0:t-1}\right) \mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right). \end{aligned} \tag{5.45}$$

as $\boldsymbol{\theta}_t$ is independent of $\mathbf{x}_{0:t-1}$ and hence $\mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \mathbf{x}_{0:t-1}, \boldsymbol{\theta}_{0:t-1}\right)$ reduces to $\mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right)$. Therefore, the MMSE estimate of the unknown variables, $\boldsymbol{\varphi}_{0:t}$, can be written as

$$\widehat{\boldsymbol{\varphi}}_t = \int_{\mathcal{X}} \int_{\Theta} \begin{bmatrix} \mathbf{x}_t \\ \boldsymbol{\theta}_t \end{bmatrix} \mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t}, \mathbf{x}_{0:t-1}\right) \mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right) \mathrm{d}\boldsymbol{\theta}_t \mathrm{d}\mathbf{x}_t \tag{5.46a}$$

$$= \begin{bmatrix} \int_{\Theta} \int_{\mathcal{X}} \mathbf{x}_t \, \mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t}, \mathbf{x}_{0:t-1}\right) \mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right) \mathrm{d}\boldsymbol{\theta}_t \mathrm{d}\mathbf{x}_t \, \mathrm{d}\boldsymbol{\theta}_t \\ \int_{\Theta} \int_{\mathcal{X}} \boldsymbol{\theta}_t \, \mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t}, \mathbf{x}_{0:t-1}\right) \mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right) \mathrm{d}\boldsymbol{\theta}_t \mathrm{d}\mathbf{x}_t \, \mathrm{d}\boldsymbol{\theta}_t. \end{bmatrix} \tag{5.46b}$$

By slightly reordering for the corresponding integrals, one thus obtains:

$$\widehat{\boldsymbol{\varphi}}_t = \begin{bmatrix} \int_{\Theta} \mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right) \left[\int_{\mathcal{X}} \mathbf{x}_t \, \mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t}, \mathbf{x}_{0:t-1}\right) \mathrm{d}\mathbf{x}_t \right] \mathrm{d}\boldsymbol{\theta}_t \\ \int_{\Theta} \boldsymbol{\theta}_t \, \mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right) \left[\int_{\mathcal{X}} \mathrm{p}\left(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t}, \mathbf{x}_{0:t-1}\right) \mathrm{d}\mathbf{x}_t \right] \mathrm{d}\boldsymbol{\theta}_t \end{bmatrix} \tag{5.46c}$$

$$= \begin{bmatrix} \int_{\Theta} \mathrm{p}\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right) \widehat{\mathbf{x}}_t \, \mathrm{d}\boldsymbol{\theta}_t \\ \widehat{\boldsymbol{\theta}}_t \end{bmatrix} = \begin{bmatrix} \widehat{\mathbf{x}}_t \mid \boldsymbol{\theta}_t \\ \widehat{\boldsymbol{\theta}}_t \end{bmatrix}. \tag{5.46d}$$

where $\widehat{\mathbf{x}}_t \mid \boldsymbol{\theta}_t$ denotes that the estimate, $\widehat{\mathbf{x}}_t$, is dependent on $\boldsymbol{\theta}_t$. Therefore, in order to obtain the MMSE of the set of unknown variables, the MMSE estimators of $\boldsymbol{\theta}_t$ and $\mathbf{x}_t$ are computed independently of each other. In other words, the optimal estimator of the unknown variables, $\boldsymbol{\varphi}_t$, *marginalises* $\mathbf{x}_t$ from $\mathrm{p}\left(\boldsymbol{\theta}_t, \mathbf{x}_t \mid \mathbf{y}_{1:t}, \mathbf{x}_t, \boldsymbol{\theta}_{0:t-1}\right)$ to obtain the

marginal posterior pdf

$$p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}) = \int_{\mathcal{M}} p(\boldsymbol{\theta}_t, \mathbf{x}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1})\, d\mathbf{x}_t \tag{5.47}$$

$$= \int_{\mathcal{M}} p(\mathbf{x}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t}, \mathbf{x}_{0:t-1})\, p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1})\, d\mathbf{x}_t \tag{5.48}$$

If one of the variables, for argument's sake $\mathbf{x}_t$, is analytically tractable, $\mathbf{x}_t$ can be estimated using its optimal estimator, whereas $\boldsymbol{\theta}_{0:t}$ are obtained using the SIR particle filtering. Again, recalling sect. §5.3 on page 83, the source signal of the state space in eqn. (5.6) on page 83 can be estimated using the Kalman filter described by the Kalman filter equations in eqn. (5.10). According to eqn. (5.46), the source signal and model parameters can be jointly estimated by drawing N importance samples, $\boldsymbol{\theta}_t^{(i)}$, of the model parameters. For each particle, the source signal at time t, $\mathbf{x}_t^{(i)}$, is evaluated using the Kalman filter equations. Using $\boldsymbol{\theta}_t^{(i)}$ and $\mathbf{x}_t^{(i)}$, the weight, $w_t^{(i)}$, is computed for each particles and the variables, $\boldsymbol{\varphi}_t$, are resampled according to their weights if the effective sample size, N_{eff}, lies below a predetermined threshold.

The resulting particle filter framework is known as the Rao-Blackwellized particle filter (RBPF) and was implemented in, e.g., [24, 203–205]. A schematic illustration is shown in Fig. 5.6. As the RBPF provides improvement in the variance of the estimates and hence facilitates more accurate estimation of the analytically tractable substructures, the proposed dereverberation framework in Chap. 6 is heavily based on Rao-Blackwellisation of the unknown parameter space.

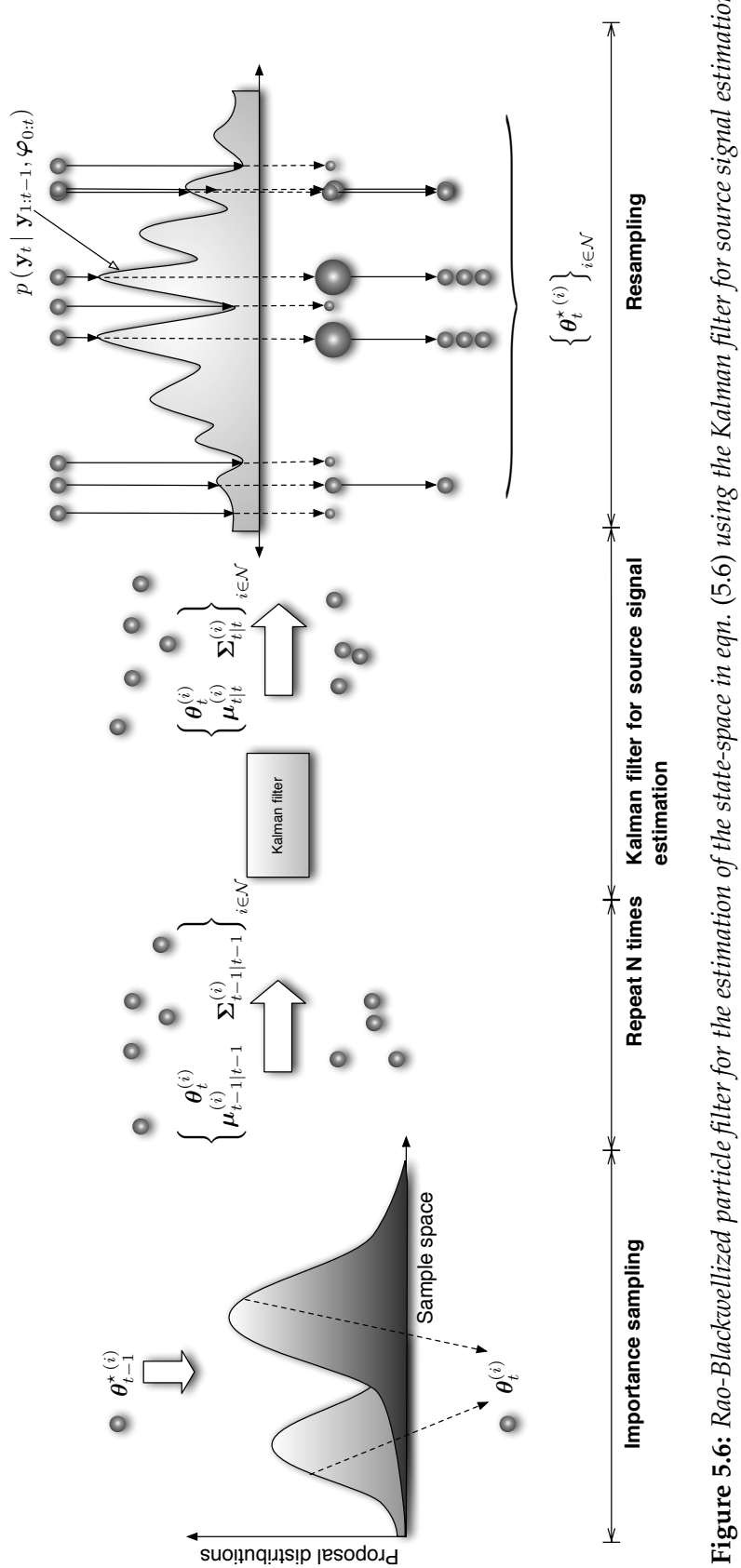


Figure 5.6: *Rao-Blackwellized particle filter for the estimation of the state-space in eqn.* (5.6) *using the Kalman filter for source signal estimation.*

5.7 Summary

This chapter discussed the methodology required in the remainder of this thesis. When performing signal estimation, the estimates are desired to be optimal in the sense of a statistical mode. ML, MAP, and MMSE are popular choices of modes, optimising either the likelihood, posterior pdf of error of the estimate with respect to the measurements and clean source signal. One of the most widely used estimators, yielding optimal estimates in the MSE sense of Gaussian state spaces, is the Kalman filter. Facilitating sequential processing, the Kalman filter is particularly well suited for online signal processing. However, Kalman filters require knowledge of the underlying model parameters, which are generally unavailable for blind dereverberation problems. The optimality of Kalman filters can be exploited, whilst avoiding the necessity of explicit knowledge of the model parameters, by incorporating the Kalman filter in a particle filter framework. Particle filters are SMC methods that obtain estimates of unknown and analytically intractable variables. Knowledge is inferred from the most immediate observations and is sequentially used for updating and tracking the parameter space. By incorporation of the Kalman filter in a particle filter, sub-spaces of smaller dimensions are evaluated separately, reducing the variance of the estimator. As the optimal estimator of a subset of the unknown variables is included, the particle filter is Rao-Blackwellised.

Using the methodology discussed in this chapter and the speech production and channel model examined in Chaps. 3 and 4, the proposed frame work for blind speech dereverberation is derived in the following chapter.

Part III

Proposed methodology

Chapter 6

Blind speech dereverberation by marginalisation of the acoustic channel

6.1 Introduction

This chapter proposes to a Rao-Blackwellised particle filter framework for blind speech dereverberation, facilitating flexible incorporation of different source models and the application to both dereverberation of speech from stationary and moving speakers.

The underlying idea of the proposed approach is that the 1) source signal, 2) channel model parameters, and 3) source model parameters and noise variance terms can be estimated using three different estimators. Recalling the discussions on source models in Chap. 3 and noise models in sect. §4.6 on page 76, the source model parameters and noise variance terms are independent of both the source signal and channel models. Therefore, the model parameters excluding the channel can be estimated directly from the observed signal independently of either the channel or source signal. Due to non-linearities of the observations in the source model parameters, an optimal estimator cannot be derived for the remaining unknown variables. Instead, these are obtained using sequential importance resampling (SIR) as discussed in sect. §5.5.

Considering the state space representing the system model discussed in Chaps. 3 and 4 and illustrated in Fig. 6.1, given an estimate of the source model parameters and variance terms, estimation of the augmented space of the source signal and channel parameters reduces to a Kalman filter. The Kalman filters for source signal and channel estimation can therefore be integrated with the estimation of the remaining model parameters in a Rao-Blackwellized particle filter (RBPF) framework (see sect. §5.6).

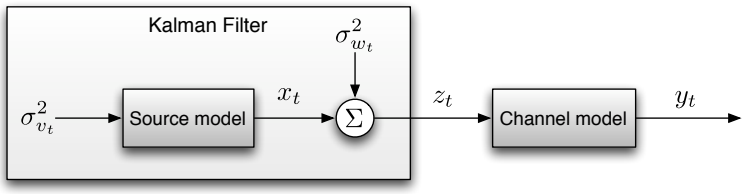


Figure 6.1: *Principle of proposed algorithm by marginalisation of channel*

In other words, the proposed algorithm can therefore be interpreted as the evaluation of an *ensemble* of Kalman filters for the source signal and channel parameters for a *stochastically selected* set of source parameters and variance terms.

This chapter is structured as follows: Sect. §6.2 introduces the system model, whilst sect. §6.3 derives the corresponding marginal probability density functions (pdfs), showing that the channel parameters and source signal, and the remaining model parameters can be estimated separately. Sect. §6.4 derives the Kalman filter for source signal and channel estimation, whilst sect. §6.5 discusses the estimation of the remaining model parameters using SIR particle filtering. Conclusions are drawn in sect. §6.6.

6.2 System model

Given a reverberant and noisy observed signal received at one or several microphones, speech dereverberation can be interpreted as the estimation of the anechoic speech signal recorded directly at the speakers mouth. Considering only the information inferred from the measured reverberant signal, estimation of the source signal could be considered from a maximum likelihood perspective as discussed in sect. §5.2.1 on page 81. However, maximum-likelihood (ML) approaches require the maximisation over the multi-dimensional and possibly infinite parameter space. This search can lead to severe dimensionality issues and computational burden.

In order to reduce the dimension of the parameter space, it can therefore prove highly advantageous to incorporate prior information about the source production mechanism and distorting channel in the estimation process. As exact knowledge of the speech production mechanism and distorting channel are generally unavailable, models of the vocal tract and room acoustics are utilised instead as discussed in Chaps. 3 and 4. Based on these discussions, sect. §6.2.2 to sect. §6.2.3 summarise the system model utilised for the development of the proposed algorithm.

6.2.1 General TVAR source model

Chap. 3 showed that the speech production mechanism can be modelled as a concatenation of lossless acoustic tubes of equal lengths, whose transfer function can be approximated by an all-pole filter. Therefore, the speech signal can be modelled as a time-varying AR (TVAR) process of order Q as demonstrated in eqn. (3.18) on page 48:

$$x_t = \sum_{q \in \mathcal{Q}} a_{q,t}\, x_{t-q} + \sigma_{v_t}\, v_t, \tag{6.1}$$

where $\{x_t\}_{t \geq Q}$ are the source signal samples for time $t \geq 0$, $\{a_{q,t}\}_{q \in \mathcal{Q}}$ are the TVAR parameters, and $v_t \sim \mathcal{N}(0, 1)$ is the process excitation with variance $\sigma^2_{v_t}$.

It is crucial to note that the source *signal* model in eqn. (6.1) only specified that the source obeys a TVAR process, but leaves the source *parameter* model, and hence the dynamic properties of the source signal, unspecified. Eqn. (6.1) therefore provides a *general* source model. Specific characteristics, such as harmonicity, can be enforced by specifying corresponding parameter models on $\{a_{q,t}\}_{q \in \mathcal{Q}}$.

6.2.2 General all-pole channel model

Whilst speech models were extensively discussed in Chap. 3, Chap. 4 reviewed the modelling of reverberant channels. It was shown that the transfer function of geometric reverberant room can be approximated by all-pole models. The observed signal at sensor $m \in \mathcal{M}$ can therefore be expressed as

$$y_{m,t} = \sum_{p \in \mathcal{P}} b_{m,p}\, y_{m,t-p} + x_t, \tag{6.2}$$

where $\{y_t\}_{t \geq P+Q+1}$ are the reverberant observed signal samples for time $t \geq P+Q+1$, P is the channel order, and $\{b_p\}_{p \in \mathcal{P}}$ are the channel parameters. For stationary speakers, the channel parameters are static and hence time invariant. In contrast, the room transfer function (RTF) varies with the source-sensor distance, and hence varies with time as the speaker is moving, i.e., changing position with time. Thus, for moving speakers, a dynamic is induced on the channel parameters $b_p = b_{p,t}$ similar to the TVAR speech parameters, such that parameter models on $b_{p,t}$ are required (see sect. §9.3 on page 175). Again, as this chapter focuses on the derivation of a *generalised* speech dereverberation algorithm, dynamic parameter models are discussed in Chap. 9 where the dereverberation of speech from a moving speaker is investigated.

Furthermore, as discussed in sect. §4.6, in scenarios where a noise source is located closely to the speaker – e.g., for conference calls from a computer workstation –, the

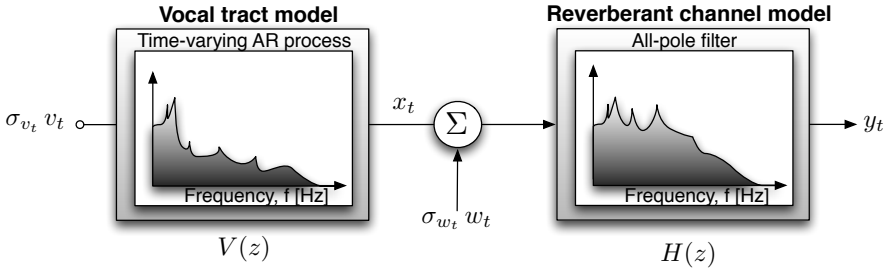


Figure 6.2: *Generation of reverberant signals using the source and channel model developed in Chaps. 3 and 4.*

source signal is first distorted by the noise and subsequently filtered with the reverberant channel, such that eqn. (6.2) can be expanded to eqn. (4.15) on page 77, i.e.,

$$y_{m,t} = \sum_{p \in \mathcal{P}} b_{m,p}\, y_{m,t-p} + x_t + \sigma_{w_{m,t}}\, w_{m,t}, \tag{6.3}$$

where $w_{m,t}$ is the noise signal with variance $\sigma^2_{w_t}$. white Gaussian noise (WGN) sources are assumed here, such that $w_{m,t} \sim \mathcal{N}(0,\, 1)$. Furthermore, time-varying properties of the interference is incorporated by letting the log-variance evolve according to a random walk similar to the process noise in eqn. (3.13) on page 42, i.e.,

$$\phi_{w_{m,t}} = \phi_{w_{m,t-1}} + \sigma_{\phi_{w_{m,t}}}\, r_{\phi_{w_{m,t}}}, \qquad r_{\phi_{w_{m,t}}} \sim \mathcal{N}(0,\, 1) \tag{6.4}$$

or, equivalently in form of a pdf:

$$p\left(\phi_{w_{m,t}} \mid \phi_{w_{m,t-1}}\right) = \mathcal{N}\left(\phi_{w_{m,t}} \,\middle|\, \phi_{w_{m,t-1}},\, \sigma^2_{\phi_{w_{m,t}}}\right) \tag{6.5}$$

Again, note that the channel model can be easily expanded to that of a moving speaker by extending b_p to a time-varying $b_{p,t}$ whose dynamic is specified by some parameter model similar to the source parameter models. Eqn. (6.3) therefore provides a *general* observation model.

Considering the unknown variables, i.e., the source signal, channel and source parameters, and noise variance terms, in the system model as random variables, the system model can be expressed in terms of pdfs. Rather than specifying multiple one-dimensional pdfs over each dimension of the source order, channel order, number of microphones, etc., the pdfs can be expressed more compactly in matrix form. Therefore, eqns. (6.1) and (6.3) should be phrased in state-space rather than scalar form.

6.2.3 System state space

The source model in eqn. (6.1) can be easily rewritten in matrix form as

$$\mathbf{x}_t = \mathbf{A}_t \mathbf{x}_{t-1} + \boldsymbol{\Sigma}_{\mathbf{v}_t} \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}\left(\mathbf{0}_{Q\times 1}, \mathbf{I}_Q\right). \tag{6.6}$$

Even though $\mathbf{b}$ is static, i.e., $p\left(\mathbf{b}_t \mid \mathbf{b}_{t-1}, \theta_{0:t}\right)$ is technically a Delta-function, it is desirable to include an expression of the channel in the state space model. Hence, the static channel can be expressed equivalent form as

$$\mathbf{b}_t = \mathbf{I}_{MP}\, \mathbf{b}_{t-1} + \mathbf{0}_{1\times MP}\, \mathbf{r}_t \qquad \mathbf{r}_t \sim \mathcal{N}\left(\mathbf{0}_{MP\times 1}, \mathbf{I}_{MP}\right). \tag{6.7}$$

Hence, as both eqns. (6.6) and (6.7) are of the form of a first-order Markov process excited by WGN, $\mathbf{x}_t$ and $\mathbf{b}$ can be *augmented* into a combined state, $\mathbf{z}_t \triangleq \begin{bmatrix}\mathbf{b}^T & \mathbf{x}_t^T\end{bmatrix}^T$, where,

$$\mathbf{z}_t = \mathbf{D}_t\, \mathbf{z}_{t-1} + \boldsymbol{\Sigma}_{\mathbf{D}_t}\, \mathbf{s}_t \qquad \mathbf{s}_t \sim \mathcal{N}\left(\mathbf{0}_{MP+Q\times 1}, \mathbf{I}_{MP+Q}\right) \tag{6.8}$$

where

$$\mathbf{D}_t \triangleq \begin{bmatrix}\mathbf{I}_{MP} & \mathbf{0}_{MP\times Q}\\ \mathbf{0}_{Q\times MP} & \mathbf{A}_t\end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\mathbf{D}_t}\, \boldsymbol{\Sigma}_{\mathbf{D}_t}^T \triangleq \begin{bmatrix}\mathbf{0}_{MP\times MP} & \mathbf{0}_{MP\times Q}\\ \mathbf{0}_{Q\times MP} & \boldsymbol{\Sigma}_{\mathbf{v}_t}\boldsymbol{\Sigma}_{\mathbf{v}_t}{}^T\end{bmatrix}.$$

Furthermore, the observations in eqn. (6.3) can be rewritten in matrix form similar to eqn. (6.6), such that the system state space model is expressed as:

$$\mathbf{z}_t = \mathbf{D}_t\, \mathbf{z}_{t-1} + \boldsymbol{\Sigma}_{\mathbf{D}_t} \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}\left(\mathbf{0}_{MP+Q\times 1}, \mathbf{I}_{MP+Q}\right), \tag{6.9a}$$

$$\mathbf{y}_t = \mathbf{H}_t\, \mathbf{z}_t + \boldsymbol{\Sigma}_{w_t} \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}\left(\mathbf{0}_{M\times 1}, \mathbf{I}_M\right), \tag{6.9b}$$

where $\mathbf{x}_t$ is a model-dependent vector of the source signal samples, $\mathbf{A}_t$ is the source transition matrix governed by the source model parameters, and $\boldsymbol{\Sigma}_{\mathbf{v}_t}$ is the covariance matrix of the WGN source excitation, $\mathbf{v}_t$. The structure of $\mathbf{x}_t$, $\mathbf{A}_t$, $\boldsymbol{\Sigma}_{\mathbf{v}_t}$ and $\mathbf{v}_t$ are defined by the underlying source parameter models and are specified in detail in Chaps. 7 and 8. $\mathbf{y}_t = \begin{bmatrix}y_{1,t} & \dots & y_{M,t}\end{bmatrix}^T$ are the M sensor observations, $\boldsymbol{\Sigma}_{w_t}\, \boldsymbol{\Sigma}_{w_t}^T \triangleq \operatorname{diag}\begin{bmatrix}\sigma^2_{w_{1,t}} & \dots & \sigma_{w^2_{M,t}}\end{bmatrix}$ is the $M \times M$ covariance matrix of the measurement noise, $\mathbf{z}_t \triangleq \begin{bmatrix}\mathbf{b}^T & \mathbf{x}_t^T\end{bmatrix}^T$ contains the source signal and channel parameters, and $\mathbf{H}_t \triangleq \begin{bmatrix}\mathbf{Y}_{t-1} & \mathbf{C}^T\end{bmatrix}$, where $\mathbf{C}^T = \mathbf{1}_{M\times 1}\, \mathbf{c}^T$ with $\mathbf{c}^T$ defined as a $1 \times Q$ source-model dependent combination

of ones and zeros retaining only the samples of $\mathbf{x}_t$ required for the generation of $\mathbf{y}_t$

$$\mathbf{Y}_{t-1} \triangleq \begin{bmatrix} \widehat{\mathbf{y}}_{1,t-1}^T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \widehat{\mathbf{y}}_{M,t-1}^T \end{bmatrix} \tag{6.10}$$

and where $\widehat{\mathbf{y}}_{m,t-1} \triangleq \begin{bmatrix} y_{m,t-1} & \cdots & y_{m,t-P} \end{bmatrix}^T \; \forall\, m \in \mathcal{M}$.

Given the system model in eqn. (6.9), the aim is to estimate the source signal and channel parameters contained in $\mathbf{z}_t \triangleq \begin{bmatrix} \mathbf{b}^T & \mathbf{x}_t^T \end{bmatrix}^T$ from $\mathbf{y}_t$ for each sample t. In practice, any remaining model parameters such as the source model parameters, measurement and process noise covariance are unknown for blind speech dereverberation problems. Therefore, the set of unknown parameters $\boldsymbol{\varphi}_t \triangleq \begin{bmatrix} \mathbf{z}_t^T, \boldsymbol{\theta}_t^T \end{bmatrix}^T$ is to be estimated, where $\boldsymbol{\theta}_t$ contains the remaining, unknown model parameters.

6.3 Rao-Blackwellisation of the source signal and channel

Recalling the discussion of Bayesian estimates in sect. §5.2, the minimum mean-square error (MMSE) estimate of $\boldsymbol{\varphi}_{0:t}$ can be evaluated via

$$\widehat{\boldsymbol{\varphi}}_{0:t} = \mathbb{E}_{p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})}[\boldsymbol{\varphi}_{0:t}] = \int_{\Phi} \boldsymbol{\varphi}_{0:t}\, p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})\, d\boldsymbol{\varphi}_{0:t}. \tag{5.5}$$

In order to identify the expected value in eqn. (5.5), the integral over the joint posterior pdf $p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$ needs to be solved simultaneously over all parameter spaces contained in $\boldsymbol{\varphi}_{0:t}$. As discussed in sect. §5.6, the variance of the direct evaluation of eqn. (5.5) can be reduced by evaluating substructures of $\boldsymbol{\varphi}_{0:t}$ of smaller dimension individually.

By application of the probability chain rule, the filtering distribution, $p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$ in eqn. (5.5) can be expressed as

$$\begin{aligned} p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t}) &= p(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})\, p(\boldsymbol{\varphi}_{0:t-1} \,|\, \mathbf{y}_{1:t}) \\ &= p(\boldsymbol{\varphi}_t \,|\, \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})\, p(\boldsymbol{\varphi}_{0:t-1} \,|\, \mathbf{y}_{0:t-1}) \end{aligned} \tag{6.11}$$

as $\boldsymbol{\varphi}_{0:t-1}$ is independent of $\mathbf{y}_t$. Thus, the joint pdf, $p(\boldsymbol{\varphi}_{0:t} \,|\, \mathbf{y}_{1:t})$ is obtained by sequentially updating the trajectory at $t-1$, $p(\boldsymbol{\varphi}_{0:t-1} \,|\, \mathbf{y}_{0:t-1})$, with the marginal pdf at t,

$p(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})$. The marginal pdf, $p(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})$, is equivalent to:

$$\begin{aligned} p(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}) &= p(\mathbf{z}_t, \boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \mathbf{z}_{0:t-1}, \boldsymbol{\theta}_{0:t-1}) \\ &= p(\mathbf{z}_t \mid \mathbf{y}_{1:t}, \mathbf{z}_{0:t-1}, \boldsymbol{\theta}_{0:t})\, p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \mathbf{z}_{0:t-1}, \boldsymbol{\theta}_{0:t-1}) \\ &= p(\mathbf{z}_t \mid \mathbf{y}_{1:t}, \mathbf{z}_{t-1}, \boldsymbol{\theta}_{0:t})\, p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}) \end{aligned} \tag{6.12}$$

where $\mathbf{z}_t$ obeys a first-order Markov chain and is hence independent of $\mathbf{z}_{0:t-2}$, and $\boldsymbol{\theta}_t$ is independent of $\mathbf{z}_t$. Recalling the discussion of Rao-Blackwellisation schemes in sect. §5.6 on page 103, eqn. (6.12) can be expressed as

$$\hat{\boldsymbol{\varphi}}_t = \int_{\mathcal{Z}} \int_{\Theta} \begin{bmatrix} \mathbf{z}_t \\ \boldsymbol{\theta}_t \end{bmatrix} p(\mathbf{z}_t \mid \mathbf{y}_{1:t}, \mathbf{z}_{t-1}, \boldsymbol{\theta}_{0:t})\, p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1})\, d\boldsymbol{\theta}_t d\mathbf{z}_t \tag{6.13}$$

$$= \begin{bmatrix} \int_{\Theta} \int_{\mathcal{Z}} \mathbf{z}_t\, p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1})\, p(\mathbf{z}_t \mid \mathbf{y}_{1:t}, \mathbf{z}_{t-1}, \boldsymbol{\theta}_{0:t})\, d\mathbf{z}_t\, d\boldsymbol{\theta}_t \\ \int_{\Theta} \int_{\mathcal{Z}} \boldsymbol{\theta}_t\, p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1})\, p(\mathbf{z}_t \mid \mathbf{y}_{1:t}, \mathbf{z}_{t-1}, \boldsymbol{\theta}_{0:t})\, d\mathbf{z}_t\, d\boldsymbol{\theta}_t \end{bmatrix} \tag{6.14}$$

$$= \begin{bmatrix} \int_{\Theta} p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1})\, \hat{\mathbf{z}}_t\, d\boldsymbol{\theta}_t \\ \hat{\boldsymbol{\theta}}_t \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{z}}_t \mid \boldsymbol{\theta}_t \\ \hat{\boldsymbol{\theta}}_t \end{bmatrix} \tag{6.15}$$

where $\hat{\boldsymbol{\varphi}}_t$, $\hat{\mathbf{z}}_t$ and $\hat{\boldsymbol{\theta}}_t$ denote the MMSE estimates of $\boldsymbol{\varphi}_t$, $\mathbf{z}_t$ and $\boldsymbol{\theta}_t$ respectively, Θ covers the region of support over the source parameters, the excitation covariance and the measurement covariance, and $\mathcal{Z}$ covers the region of support of the channel parameters, $\mathcal{B}_P$, and the source signal, $\mathcal{X}$. Note that $\hat{\mathbf{z}}_t|\hat{\boldsymbol{\theta}}_t$ is the estimate of the source signal and channel parameters *conditional* on $\boldsymbol{\theta}_t$. Therefore, according to eqn. (6.15), $\mathbf{z}_t$, can be marginalised from $\boldsymbol{\theta}_t$, such that $\boldsymbol{\theta}_t$ can be estimated independently of $\mathbf{z}_t$. As the source signal is dependent on knowledge of the model parameters, an estimate of $\mathbf{z}_t$ can only be obtained conditional on $\boldsymbol{\theta}_t$. Assuming that an estimate of $\hat{\boldsymbol{\theta}}_t$ is available, an estimate conditional on $\boldsymbol{\theta}_t$ of $\mathbf{z}_t$ can be obtained.

6.4 Kalman filter for source signal and channel estimation

The discussion in sect. §5.3 on page 83 highlighted that the optimal estimator of a hidden system state is a Kalman filter if the system is of the form of the conditionally Gaussian state-space (CGSS) in eqn. (5.6), i.e.,

$$\mathbf{x}_t = \mathbf{A}_t \mathbf{x}_{t-1} + \boldsymbol{\Sigma}_{\mathbf{v}_t} \mathbf{v}_t \qquad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}_{Q\times 1}, \mathbf{I}_Q) \tag{6.16a}$$

$$\mathbf{y}_t = \mathbf{C}_t \mathbf{x}_t + \boldsymbol{\Sigma}_{w_t} \mathbf{w}_t \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}_{M\times 1}, \mathbf{I}_M)\,. \tag{6.16b}$$

Comparing to the state space in eqn. (6.9), the source signal, $\mathbf{x}_t$, can hence be optimally estimated using the Kalman filter by replacing $\mathbf{C}_t$ and $\mathbf{x}_t$ in eqn. (6.16b) by $\mathbf{H}_t$ and $\mathbf{z}_t$ respectively. Therefore, the augmented state vector, $\mathbf{z}_t$, containing the source signal and channel parameters is *predicted* and *corrected* according to eqns. (5.7) and (5.8) on page 83 and on page 84 respectively, i.e.,

$$p(\mathbf{z}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}) = \mathcal{N}\left(\mathbf{z}_t \,\middle|\, \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}\right) \tag{6.17a}$$

$$p(\mathbf{z}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t}) = \mathcal{N}\left(\mathbf{z}_t \,\middle|\, \boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}\right) \tag{6.17b}$$

where the predicted and corrected states, $\boldsymbol{\mu}_{t|t-1}$ and $\boldsymbol{\mu}_{t|t}$ respectively, and their corresponding error covariance terms, $\boldsymbol{\Sigma}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t}$, are given by:

$$\boldsymbol{\mu}_{t|t-1} = \mathbf{D}_t \, \boldsymbol{\mu}_{t-1|t-1} \tag{6.18a}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{\Sigma}_{\mathbf{D}_t} \, \boldsymbol{\Sigma}_{\mathbf{D}_t}^{\mathsf{T}} + \mathbf{D}_t^{\mathsf{T}} \, \boldsymbol{\Sigma}_{t-1|t-1} \, \mathbf{D}_t \tag{6.18b}$$

$$\boldsymbol{\mu}_{t|t} = (\mathbf{I}_{MP+Q} - \mathbf{K}_t \, \mathbf{H}_t) \, \boldsymbol{\mu}_{t|t-1} - \mathbf{K}_t \, \mathbf{y}_t \tag{6.18c}$$

$$\boldsymbol{\Sigma}_{t|t} = (\mathbf{I}_{MP+Q} - \mathbf{K}_t \, \mathbf{H}_t) \, \boldsymbol{\Sigma}_{t|t-1}, \tag{6.18d}$$

Furthermore, the Kalman gain, $\mathbf{K}_t$, and residual covariance, $\boldsymbol{\Sigma}_{z_t}$, are defined as:

$$\mathbf{K}_t \triangleq \boldsymbol{\Sigma}_{t|t-1} \mathbf{H}_t^{\mathsf{T}} \boldsymbol{\Sigma}_{z_t}^{-1} \tag{6.19a}$$

$$\boldsymbol{\Sigma}_{z_t} \triangleq \mathbf{H}_t \, \boldsymbol{\Sigma}_{t|t-1} \, \mathbf{H}_t^{\mathsf{T}} + \boldsymbol{\Sigma}_{w_t} \, \boldsymbol{\Sigma}_{w_t}^{\mathsf{T}} \tag{6.19b}$$

As the Kalman filter *jointly* estimates the channel and source signal conditional on $\boldsymbol{\theta}_{0:t}$, information about the structure of the estimates in eqn. (6.18) is desired in order to determine whether *marginal*, individual estimates of the channel and source signal respectively can be extracted from their joint estimator.

6.4.1 From joint to marginal estimation

In order to determine the structure of the Kalman filter equations in eqn. (6.18), assume that the Kalman filter equations are initialised with an augmented state $\boldsymbol{\mu}_{P|P} \triangleq \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{b},P}^{\mathsf{T}} & \boldsymbol{\mu}_{\mathbf{x}_{P|P-1}}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$ and block-diagonal covariance matrix:

$$\boldsymbol{\Sigma}_{P|P} \triangleq \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{b},P} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \boldsymbol{\Sigma}_{\mathbf{x}_{P|P-1}} \end{bmatrix}$$

where $\boldsymbol{\Sigma}_{\mathbf{b},P}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_{P|P-1}}$ are the initial estimates of the channel and source signal respectively. This is a reasonable assumption as the states themselves are defined as the

augmentation of the channel and source signal, $\mathbf{z}_t \triangleq \begin{bmatrix}\mathbf{b}^T & \mathbf{x}_t^T\end{bmatrix}^T$.

Under the assumption of block-separability of $\boldsymbol{\mu}_{P|P}$ and $\boldsymbol{\Sigma}_{P|P}$, it can be shown by induction (see Appendix C.1) that the states at time t are expressed as:

$$\boldsymbol{\mu}_{t|t-1} \triangleq \begin{bmatrix}\boldsymbol{\mu}_{\mathbf{b},t-1} \\ \boldsymbol{\mu}_{\mathbf{x}_{t|t-1}}\end{bmatrix} \qquad \boldsymbol{\Sigma}_{t|t-1} \triangleq \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{b},t-1} & \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{t-1}}\,\mathbf{A}_t \\ \mathbf{A}_t^T\,\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{t-1}} & \boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}}\end{bmatrix} \tag{6.20a}$$

$$\boldsymbol{\mu}_{t|t} \triangleq \begin{bmatrix}\boldsymbol{\mu}_{\mathbf{b},t} \\ \boldsymbol{\mu}_{\mathbf{x}_{t|t}}\end{bmatrix} \qquad \boldsymbol{\Sigma}_{t|t} \triangleq \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{b},t} & \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_t} \\ \boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_t} & \boldsymbol{\Sigma}_{\mathbf{x}_{t|t}}\end{bmatrix} \tag{6.20b}$$

where $\boldsymbol{\mu}_{\mathbf{b},t}$ and $\boldsymbol{\Sigma}_{\mathbf{b},t}$ are the corrected Kalman states and covariance of the channel, $\boldsymbol{\mu}_{\mathbf{x}_{t|t-1}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}}$ are the predicted states and covariance of the source signal, and $\boldsymbol{\mu}_{\mathbf{x}_{t|t}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_{t|t}}$ are the updated states and covariance of the source signal. Furthermore, $\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_t}$ and $\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_t}$ are the cross-correlations between the channel and source.

As both the predicted and updated states and covariance terms can be separated into blocks corresponding to the channel and source signal, the *marginal* estimators of $\mathbf{b}$ and $\mathbf{x}_t$ can be expressed individually. Therefore, the source signal is estimated, conditional on $\theta_{0:t}$, by the set of equations:

$$\boldsymbol{\mu}_{\mathbf{x}_{t|t-1}} = \mathbf{A}_t\,\boldsymbol{\mu}_{\mathbf{x}_{t-1|t-1}} \tag{6.21a}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}} = \mathbf{A}_t^T\boldsymbol{\Sigma}_{\mathbf{x}_{t-1|t-1}}\mathbf{A}_t + \boldsymbol{\Sigma}_{\mathbf{v}_t}\,\boldsymbol{\Sigma}_{\mathbf{v}_t}{}^T \tag{6.21b}$$

$$\boldsymbol{\mu}_{\mathbf{x}_{t|t}} = \left(\mathbf{I}_Q - \mathbf{K}_{\mathbf{x}_t}\,\mathbf{C}^T\right)\boldsymbol{\mu}_{\mathbf{x}_{t|t-1}} + \mathbf{K}_{\mathbf{x}_t}\left(\mathbf{y}_t - \mathbf{Y}_{t-1}\,\boldsymbol{\mu}_{\mathbf{b},t-1}\right) \tag{6.21c}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}_{t|t}} = \left(\mathbf{I}_Q - \mathbf{K}_{\mathbf{x}_t}\,\mathbf{C}^T\right)\boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}} - \mathbf{K}_{\mathbf{x}_t}\,\mathbf{Y}_{t-1}\,\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{t-1}}\,\mathbf{A}_t \tag{6.21d}$$

whilst the channel is estimated, conditional on $\theta_{0:t}$, using:

$$\boldsymbol{\mu}_{\mathbf{b},t} = \left(\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_t}\,\mathbf{Y}_{t-1}\right)\boldsymbol{\mu}_{\mathbf{b},t-1} + \mathbf{K}_{\mathbf{b}_t}\left(\mathbf{y}_t - \mathbf{C}^T\,\boldsymbol{\mu}_{\mathbf{x}_{t|t-1}}\right) \tag{6.22a}$$

$$\boldsymbol{\Sigma}_{\mathbf{b},t} = \left(\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_t}\,\mathbf{Y}_{t-1}\right)\boldsymbol{\Sigma}_{\mathbf{b},t-1} - \mathbf{K}_{\mathbf{b}_t}\,\mathbf{C}^T\,\mathbf{A}_t^T\,\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{t-1}} \tag{6.22b}$$

where the Kalman gain terms of the channel, $\mathbf{K}_{\mathbf{b}_t}$, the source signal, $\mathbf{K}_{\mathbf{x}_t}$, and residual

covariance are defined as

$$\mathbf{K}_{\mathbf{b}_t} \triangleq \left(\boldsymbol{\Sigma}_{\mathbf{b},t-1}\,\mathbf{Y}_{t-1}^{\mathsf{T}} + \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{t-1}}\,\mathbf{A}_t\,\mathbf{C}\right)\boldsymbol{\Sigma}_{z_t}^{-1} \tag{6.23}$$

$$\mathbf{K}_{\mathbf{x}_t} \triangleq \left(\mathbf{A}_t^{\mathsf{T}}\,\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{t-1}}\,\mathbf{Y}_{t-1}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}}\,\mathbf{C}\right)\boldsymbol{\Sigma}_{z_t}^{-1} \tag{6.24}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{z_t} = {} & \mathbf{Y}_{t-1}\,\boldsymbol{\Sigma}_{\mathbf{b},t-1}\,\mathbf{Y}_{t-1}^{\mathsf{T}} + \mathbf{C}^{\mathsf{T}}\,\boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}}\,\mathbf{C} + \boldsymbol{\Sigma}_{w_t}\,\boldsymbol{\Sigma}_{w_t}^{\mathsf{T}} \\ & + \mathbf{C}^{\mathsf{T}}\,\mathbf{A}_t^{\mathsf{T}}\,\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{t-1}}\,\mathbf{Y}_{t-1}^{\mathsf{T}} + \mathbf{Y}_{t-1}\,\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{t-1}}\,\mathbf{A}_t\,\mathbf{C}\end{aligned} \tag{6.25}$$

and the cross-correlation terms are given by

$$\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_t} = (\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_t}\,\mathbf{Y}_{t-1})\,\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{t-1}}\,\mathbf{A}_t - \mathbf{K}_{\mathbf{b}_t}\,\mathbf{C}^{\mathsf{T}}\,\boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}} \tag{6.26}$$

$$\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_t} = \left(\mathbf{I}_Q - \mathbf{K}_{\mathbf{x}_t}\,\mathbf{C}^{\mathsf{T}}\right)\mathbf{A}_t^{\mathsf{T}}\,\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{t-1}} - \mathbf{K}_{\mathbf{x}_t}\,\mathbf{Y}_{t-1}\,\boldsymbol{\Sigma}_{\mathbf{b},t-1} \tag{6.27}$$

Comparing to the standard Kalman equations in eqn. (5.10) on page 84, the source signal is thus propagated in time using the Kalman prediction and update equations, whilst the channel is estimated using the Kalman update equations.

As the MMSE estimate of the source signal and channel, $\mathbf{z}_t$, is conditional on the remaining model parameters, $\boldsymbol{\theta}_{0:t}$, according to eqn. (6.15), an estimator of $\boldsymbol{\theta}_{0:t}$ is required to perform the Kalman filter estimation of $\mathbf{z}_t$.

6.5 Estimation of the intractable model parameters

In order to evaluate the MMSE estimate of the parameters, the expectation over $\boldsymbol{\theta}_t$ needs to be solved, i.e.,

$$\widehat{\boldsymbol{\theta}}_t = \int_{\Theta} \boldsymbol{\theta}_t\, p\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right) d\boldsymbol{\theta}_t, \tag{6.28}$$

where the posterior pdf of the model parameters can be expressed as

$$p\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t-1}\right) = \frac{p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}\right) p\left(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t-1}\right)}{p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t-1}\right)}. \tag{6.29}$$

The marginal likelihood, $p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}\right)$, in eqn. (6.29) is obtained by marginalising $\mathbf{z}_t$ from the likelihood that can be straightforwardly obtained by probability transformation of eqn. (6.9b) on page 115. By slightly rewriting eqn. (6.9b) as:

$$\boldsymbol{\Sigma}_{w_t}\,\mathbf{w}_t = \mathbf{y}_t - \mathbf{H}_t\,\mathbf{z}_t \qquad\qquad \mathbf{w}_t \sim \mathcal{N}\left(\mathbf{0}_{MP\times 1},\, \mathbf{I}_{MP}\right)$$

and applying the probability transformation from $\mathbf{w} \rightarrow \mathbf{y}_t$, the likelihood becomes:

$$p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \mathbf{z}_t, \boldsymbol{\theta}_{0:t}\right) = \mathcal{N}\left(\mathbf{y}_t \,\middle|\, \mathbf{H}_t\,\mathbf{z}_t,\, \boldsymbol{\Sigma}_{w_t}\,\boldsymbol{\Sigma}_{w_t}^{\top}\right), \tag{6.30}$$

such that the *marginal* likelihood is obtained via integration of $\mathbf{z}_t$, i.e.,

$$p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}\right) = \int_{\mathcal{Z}} p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \mathbf{z}_t, \boldsymbol{\theta}_{0:t}\right)\, p\left(\mathbf{z}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}\right) d\mathbf{z}_t, \tag{6.31}$$

which is equivalent to (see Appendix C.2):

$$p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}\right) = \mathcal{N}\left(\mathbf{y}_t \,\middle|\, \mathbf{H}_t\boldsymbol{\mu}_{t|t-1},\, \boldsymbol{\Sigma}_{z_t}\right) \tag{6.32}$$

The evidence term, $p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}\right)$, in eqn. (6.29) is obtained by marginalising $\boldsymbol{\theta}_{0:t}$ from the marginal likelihood, $p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}\right)$, i.e.,

$$p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}\right) = \int_{\Theta} p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}\right) p\left(\boldsymbol{\theta}_{0:t} \mid \mathbf{y}_{1:t-1}\right) d\boldsymbol{\theta}_{0:t} \tag{6.33}$$

However, the likelihood, $p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}\right)$ is generally non-linear in $\boldsymbol{\theta}_{0:t}$ due to the form of the source transition model $\mathbf{A}_t$. Hence, eqn. (6.33) is analytically intractable and a closed-form solution for eqn. (6.29) cannot be found. As the optimal estimator of the model parameters cannot be expressed in closed form, approximate estimation techniques, such as Monte Carlo integration as described in sect. §5.4, therefore have to be utilised instead to obtain estimates of $\boldsymbol{\theta}_t$ as discussed in sect. §6.5.

Theoretically, an *ensemble* of Kalman filters could be evaluated over a deterministic grid of all possible permutations of parameter choices, $\boldsymbol{\theta}_t$, within the region of support Φ. By evaluating the corresponding likelihood function, $p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}\right)$, in eqn. (6.32), the ML estimate over the resulting grid of Kalman filters can be obtained.

However, from a practical perspective, this deterministic approach requires the evaluation of an infinite grid of Kalman filters and is hence computationally inefficient. In a more practical approach, an ensemble of Kalman filters can be evaluated for *stochastically sampled* parameter choices.

6.5.1 SMC approach to parameter estimation

Rather than attempting to optimise the analytically intractable model parameters based on the observed signal only over a parameter space of infinite dimension, prior information about the parameters can be taken into account in order to reduce the dimension of the parameter space to be searched.

Particle filters, as introduced in sect. §5.5, stochastically sample parameter choices from a hypothesis distribution reflecting prior knowledge about the posterior pdf. Therefore, rather than trying to directly draw estimates of the parameters directly from the intractable posterior pdf, a large number of samples (or particles) are drawn from the hypothesis distribution describing any subjective *belief* about the posterior pdf:

$$\theta_t^{(i)} \sim \pi\left(\theta_t \,|\, \mathbf{y}_{1:t}, \theta_{0:t-1}\right), \tag{6.34}$$

where $\{\theta_t\}_{i\in\mathcal{N}}$ are the N particles and $\pi\left(\theta_t \,|\, \mathbf{y}_{1:t}, \theta_{0:t-1}\right)$ is the proposal distribution.

In order to correct for any discrepancy between the posterior pdf and the hypothesis distribution, the measured data is taken into consideration by means of the importance weights associated with each particle (see eqn. (5.29) on page 94), i.e.,

$$w_t^{\star\,(i)} = \frac{p\left(\theta_t^{(i)} \,\middle|\, \mathbf{y}_{1:t}, \theta_{0:t-1}^{(i)}\right)}{\pi\left(\theta_t^{(i)} \,|\, \mathbf{y}_{1:t}, \theta_{0:t-1}^{(i)}\right)} = \frac{p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t}^{(i)}\right) p\left(\theta_t^{(i)} \,\middle|\, \mathbf{y}_{0:t-1}, \theta_{0:t-1}^{(i)}\right)}{p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t-1}^{(i)}\right) \pi\left(\theta_t^{(i)} \,|\, \mathbf{y}_{1:t}, \theta_{0:t-1}^{(i)}\right)} \tag{6.35}$$

$$= \frac{p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t}^{(i)}\right) p\left(\theta_t^{(i)} \,\middle|\, \theta_{0:t-1}^{(i)}\right)}{p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t-1}^{(i)}\right) \pi\left(\theta_t^{(i)} \,|\, \mathbf{y}_{1:t}, \theta_{0:t-1}^{(i)}\right)} \tag{6.36}$$

where $w_t^{\star\,(i)}\ \forall\ i \in \mathcal{N}$ denote the importance weights, $p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t}, \theta_{0:t}\right)$ is the likelihood given by eqn. (6.32) on page 121, $p\left(\theta_t \,|\, \theta_{0:t-1}\right)$ is the prior pdf of the parameters, and $p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t-1}\right)$ is the normalising evidence term, independent of θ_t.

The evidence, $p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t-1}\right)$, is not available in closed form as discussed in sect. §6.3 on page 116. Furthermore, as $p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t-1}\right)$ is independent of θ_t, it acts as a normalising factor only. Therefore, to circumvent analytical non-tractability of the evidence, the discrepancy between the posterior pdf and the hypothesis distribution can be evaluated using *unnormalised* weights, $w_t^{(i)}$, instead, where

$$w_t^{(i)} = \frac{p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t}^{(i)}\right) p\left(\theta_t^{(i)} \,\middle|\, \theta_{0:t-1}^{(i)}\right)}{\pi\left(\theta_t^{(i)} \,|\, \mathbf{y}_{1:t}, \theta_{0:t-1}^{(i)}\right)} \tag{6.37}$$

which, using prior importance sampling (see sect. §5.5.2.1 on page 97), simplifies to:

$$w_t^{(i)} = p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t}^{(i)}\right). \tag{6.38}$$

Note that eqn. (6.37) is equivalent to eqn. (6.35) disregarding the evidence term in the

for $t > \max\{P, Q\}$ **do**
  for $i = 1, \ldots, N$ **do**
1 Importance sampling of $\theta_{0:t}^{(i)}$;
2 Kalman filter prediction of $\boldsymbol{\mu}_{t|t-1}^{(i)}$, $\boldsymbol{\Sigma}_{t|t-1}^{(i)}$ (eqns. (6.21a) and (6.21b)).;
3 Kalman filter estimation of $\boldsymbol{\mu}_{\mathbf{b},t}^{(i)}$ and $\boldsymbol{\Sigma}_{\mathbf{b},t}^{(i)}$ (eqn. (6.22));
4 Kalman filter correction of $\boldsymbol{\mu}_{t|t}^{(i)}$, $\boldsymbol{\Sigma}_{t|t}^{(i)}$ (eqns. (6.21c) and (6.21d));
5 Evaluation of weights $w_t^{(i)}$ (eqns. (6.38), (6.32));
  end
6 Normalization of importance weights;
7 Resampling;
8 Computation of particle average:

$$\widehat{\mathbf{x}}_t = \sum_{i \in \mathcal{N}} \hat{\boldsymbol{\mu}}_{t|t}^{(i)} \qquad \widehat{\boldsymbol{\theta}}_t = \sum_{i \in \mathcal{N}} \theta_{0:t}^{(i)}, \qquad \widehat{\mathbf{b}} = \sum_{i \in \mathcal{N}} \boldsymbol{\mu}_{\mathbf{b},t}^{(i)}.$$

end

Algorithm 6.1: RBPF

denominator. The weights are re-normalised over their sum of all particles, such that

$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i \in \mathcal{N}} w_t^{(i)}}. \tag{6.39}$$

The integral of the MMSE estimate of the parameters in eqn. (6.28) on page 120 can thus be approximated by the discrete sum in eqn. (5.33) on page 95, i.e.,

$$\hat{\boldsymbol{\theta}}_t = \sum_{i \in \mathcal{N}} \boldsymbol{\theta}_t^{(i)} \tilde{w}_t^{(i)}, \tag{6.40}$$

In order to avoid particle depletion and retain statistically relevant samples only, resampling schemes are utilised if the effective sample size of the particles falls below a certain threshold as discussed in sect. §5.5.3 on page 98.

To summarise, estimates of the parameters are obtained by sequential importance sampling (SIS). Consequently, the source signal and channel are estimated using their optimal Kalman estimator in a Rao-Blackwellized particle filter (RBPF) framework. In other words, an ensemble of the Kalman filter for source signal and channel estimation is evaluated for stochastically sampled model parameter choices. The resulting likelihoods are used to evaluate the weights of the particles according to eqn. (6.37). The overall estimate of the source signal, channel and remaining model parameters corresponds to the respective particle averages. Finally, if the effective sample size is below a certain threshold, the particles are resampled according to their normalised importance weights in eqn. (6.39). The resulting RBPF is summarised in Alg. 6.1.

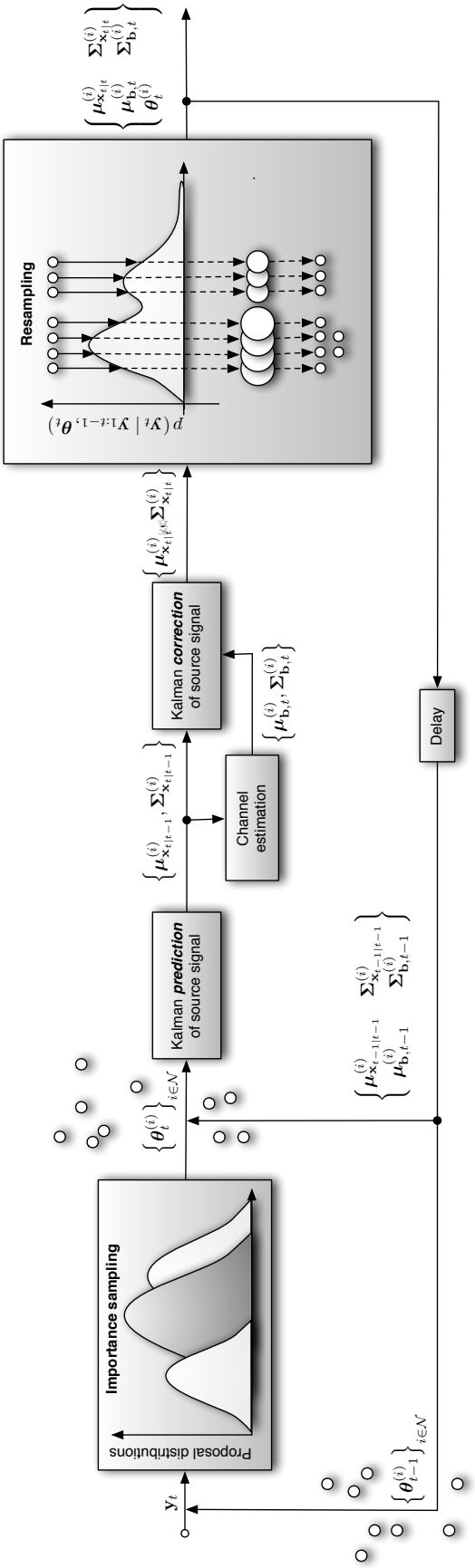


Figure 6.3: *Rao-Blackwellized particle filter*

6.6 Discussion

This chapter proposed a novel approach for signal enhancement from reverberant channels modelled as infinite impulse response (IIR) filters and noise. The vocal production system was modelled by a time-varying AR source model. The source is distorted by close-by WGN noise sources and filtered by an acoustic channel modelled by a stationary all-pole filter. Based on this system model, the source signal and channel parameters can be marginalised from the remaining model parameters and estimated using their optimal estimator, the Kalman filter. As the model parameters are unknown in practice, the Kalman filters for source and channel estimation are evaluated for a cloud of stochastically samples parameter choices using a Rao-Blackwellized particle filter (RBPF). By evaluating the likelihood of the resulting estimates, only statistically relevant estimates are retained and carried on to the next recursion. The elimination of irrelevant samples is important as the recursion at t depends on the estimation results at $t-1$ and the propagation of irrelevant samples would thus lead to estimate degeneracy.

As analytically tractable substructures of the system are obtained using their optimal estimator – the Kalman filter – the variance of the estimates is intuitively decreased [26] as compared to estimating all unknowns using a single estimator. Rao-Blackwellized particle filters marginalising the source signal by means of Kalman filter estimation from the remaining unknowns are well known in the literature (see, e.g., [132, 206]). The novelty of the approach proposed in this chapter therefore lies in its novel application to systems distorted by IIR channel models. As particle filters track and update the evolution of state and parameter trajectories with time, a dynamic is implicitly enforced on the estimated variables. Although particle filters are particularly well suited for tracking dynamic variables, such as the channel parameters of a moving speaker, the inclusion of *static* channel parameters for stationary speakers leads to particle impoverishment. The issue of static parameter estimation in particle filters is circumvented in the approach proposed in this chapter by analytically marginalising the channel from the source signal in the Rao-Blackwellized particle filter framework [132].

Furthermore, the proposed approach circumvents several problems encountered with approaches, i.e.:

1. *direct source signal estimation* rendering speech synthesis unnecessary as encountered in linear predictive coding (LPC) approaches as discussed in sect. §2.5 on page 23. Natural artefacts in the anechoic speech signal can thus be retained in the estimated signal. Furthermore, inverse filtering with an estimated equalis-

ing filter, as is the case for many explicit source modelling approaches as discussed in sect. §2.4 on page 21, does not need to be performed, circumventing non-minimum phase problems or scaling of errors;
2. *Sequential processing* facilitating real-time speech enhancement; and
3. *Blind channel estimation*, i.e., no prior knowledge of the room impulse response (RIR) is necessary as opposed to many spectral enhancement techniques (see sect. §2.3).

The performance of the RBPF is investigated and discussed in the following chapter for synthetic and real signals.

Chapter 7

Dynamic TVAR parameter model for unvoiced speech

7.1 Introduction

Chap. 6 derived the RBPF in a general form that allows for blind speech dereverberation using any speech model facilitating a CGSS representation. In order to apply the proposed RBPF particle to speech data, appropriate speech models thus need to be investigated.

As discussed in sect. §3.4.1 on page 49, the parameters of speech vary relatively smoothly and can be approximated by a random walk. Therefore, this chapter implements the RBPF for the dynamic TVAR parameter model. Results based on synthetic and speech data for a single and multiple sensors are demonstrated on synthetic channels as well as baseband RIRs.

This chapter is therefore structured as follows: Sect. §7.2 discusses the proposes parameter model. Sect. §7.3 demonstrates how the RBPF works for the proposed model. Sect. §7.4 presents results for synthetic and speech data, demonstrating that the quality of speech can be significantly improved using the dynamic TVAR parameter model for the RBPF over the reverberant signal. Conclusions are drawn in sect. §7.5.

7.2 Source model

As discussed in sect. §3.4.1, the TVAR parameters extracted from a real speech sequence modelled as the TVAR process in eqn. (3.18) on page 48, i.e.,

$$x_t = \sum_{q \in \mathcal{Q}} a_{q,t} x_{t-q} + \sigma_{v_t} v_t \qquad\qquad v_t \sim \mathcal{N}(0, 1)\,. \tag{7.1}$$

vary relatively slowly and smoothly with time as shown in Fig. 3.10a on page 47. According to eqn. (3.19) on page 49, the smooth and slowly varying behaviour can be represented by a first-order Markov chain with low variance on the parameters, $\mathbf{a}_t = \begin{bmatrix} a_{1,t} & \dots & a_{Q,t} \end{bmatrix}^T$, i.e., [130–133]

$$\mathbf{a}_t = \mathbf{a}_{t-1} + \boldsymbol{\Sigma}_{\mathbf{a}_t}\, \mathbf{r}_{\mathbf{a}_t} \qquad\qquad \mathbf{r}_{\mathbf{a}_t} \sim \mathcal{N}(\mathbf{0}_{Q\times 1}, \mathbf{I}_Q) \tag{7.2}$$

such that the prior pdf is expressed as eqn. (3.21) on page 50, i.e.,

$$p(\mathbf{a}_t \mid \mathbf{a}_{t-1}) = \mathcal{N}\left(\mathbf{a}_t \,\middle|\, \mathbf{a}_{t-1}, \boldsymbol{\Sigma}_{\mathbf{a}_t}\right). \tag{7.3}$$

It would therefore stand to reason that the dynamic TVAR parameter model in eqn. (7.2) approximates the TVAR parameters extracted from speech appropriately. Therefore, in order to incorporate the parameter model in the CGSS utilised for the RBPF, eqns. (7.1) and (7.2) are to be written in the form of eqn. (6.9a). This can be easily done by letting

$$\mathbf{A}_t = \begin{bmatrix} & \mathbf{a}_t^T & \\ \mathbf{I}_{Q-1} & & \mathbf{0}_{Q-1\times 1} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\mathbf{v}_t} = \begin{bmatrix} \sigma_{v_t}^2 & 0 & \dots & 0 \\ \vdots & & \ddots & 0 \\ 0 & 0 & \dots & 0 \end{bmatrix} \tag{7.4}$$

in the state space in eqn. (6.9a) on page 115:

$$\mathbf{x}_t = \mathbf{A}_t \mathbf{x}_{t-1} + \boldsymbol{\Sigma}_{\mathbf{v}_t} \mathbf{v}_t, \qquad\qquad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}_{Q\times 1}, \mathbf{I}_Q) \tag{7.5}$$

where $\mathbf{x}_t = \begin{bmatrix} x_t & \dots & x_{t-Q+1} \end{bmatrix}^T$ contains the Q most current source signal samples and, as discussed in sect. §3.3.2 on page 42 the process covariance noise varies according to the random walk and prior pdf in eqns. (3.13) and (3.14) respectively:

$$\phi_{v_t} = \phi_{v_{t-1}} + \sigma_{\phi_{v_t}}\, r_{\phi_{v_t}}, \qquad\qquad r_{\phi_{v_t}} \sim \mathcal{N}(0, 1) \tag{7.6}$$

$$p\left(\phi_{v_t} \mid \phi_{v_{t-1}}\right) = \mathcal{N}\left(\phi_{v_t} \,\middle|\, \phi_{v_{t-1}}, \sigma_{\phi_{v_t}}^2\right). \tag{7.7}$$

where, again, $\phi_{v_t} \triangleq \ln \sigma_{v_t}^2$ is the logarithmic value of the excitation variance, $\sigma_{v_t}^2$, and $\sigma_{\phi_{v_t}}$ is assumed constant and known.

The unknown variables in the system are therefore the source signal, $\mathbf{x}_t$, the channel model, $\mathbf{b}$, as well as the time-varying model parameters and covariance terms of the process excitation, ϕ_{ν_t}, and the M sensor noises, $\boldsymbol{\phi}_{\mathbf{w}_t}$, i.e., $\theta_t \triangleq \begin{bmatrix} \mathbf{a}_t & \phi_{\nu_t} & \boldsymbol{\phi}_{\mathbf{w}_t} \end{bmatrix}^T$. Recalling the RBPF summarised in Alg. 6.1, the source signal and channel are obtained using their optimal estimators, whereas θ_t is acquired using SIR.

7.2.1 Importance sampling of the time-varying model parameters

For optimal estimation of the time-varying model parameters, θ_t, it would be desirable to utilise the optimal importance sampling function in eqn. (5.35) on page 96, i.e., [186]

$$\pi\left(\theta_t \,|\, \mathbf{y}_{1:t}, \theta_{0:t-1}\right) = p\left(\theta_t \,|\, \mathbf{y}_{1:t}, \theta_{0:t-1}\right). \tag{7.8}$$

As discussed in sect. §5.5.2 on page 96, the optimal importance weights are obtained via eqn. (5.36), i.e.,

$$\begin{aligned} w_{0:t} &= w_{0:t-1} \times p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t-1}\right) \\ &= w_{0:t-1} \times \int p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\theta_t \,|\, \theta_{0:t-1}\right) \mathrm{d}\theta_t \end{aligned} \tag{7.9}$$

using eqn. (5.37) on page 97. As $\theta_t \triangleq \begin{bmatrix} \mathbf{a}_t & \phi_{\nu_t} & \boldsymbol{\phi}_{\mathbf{w}_t} \end{bmatrix}^T$, the prior pdf, $p\left(\theta_t \,|\, \theta_{0:t-1}\right) = \begin{bmatrix} p\left(\mathbf{a}_t \,|\, \mathbf{a}_{t-1}\right) & p\left(\phi_{\nu_t} \,|\, \phi_{\nu_{t-1}}\right) & p\left(\boldsymbol{\phi}_{\mathbf{w}_t} \,|\, \boldsymbol{\phi}_{\mathbf{w}_{t-1}}\right) \end{bmatrix}^T$, such that

$$w_{0:t} = w_{0:t-1} \times \iiint p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t}\right) \begin{bmatrix} p\left(\mathbf{a}_t \,|\, \mathbf{a}_{t-1}\right) \\ p\left(\phi_{\nu_t} \,|\, \phi_{\nu_{t-1}}\right) \\ p\left(\boldsymbol{\phi}_{\mathbf{w}_t} \,|\, \boldsymbol{\phi}_{\mathbf{w}_{t-1}}\right) \end{bmatrix} \mathrm{d}\mathbf{a}_t \, \mathrm{d}\phi_{\nu_t} \, \mathrm{d}\boldsymbol{\phi}_{\mathbf{w}_t} \tag{7.10}$$

Recalling the likelihood function, $p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t}\right)$ from eqn. (5.9) on page 84:

$$p\left(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_{0:t}\right) = \mathcal{N}\left(\mathbf{y}_t \,\middle|\, \mathbf{H}_t \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{z_t}\right), \tag{7.11}$$

it is observed that the likelihood function is related to the source parameters, $\mathbf{a}_t$, via $\mathbf{A}_t$ in $\boldsymbol{\mu}_{t|t-1}$. However, due to the form of the transition matrix in eqn. (7.4), the likelihood is, in fact, non-linear in the source parameters, such that the integral over $\mathbf{a}_t$ in eqn. (7.10) is non-tractable. Thus, the optimal importance weights cannot be evaluated analytically and optimal importance sampling cannot be used to obtain θ_t.

Instead, prior importance sampling as discussed in sect. §5.5.2.1 is used. The prior pdfs of the time-varying model parameters contained in θ_t are given by eqn. (7.3)

for $t > \max\{P, Q\}$ **do**
  for $i = 1, \ldots, N$ **do**
1 Prior importance sampling of $\theta_t^{(i)}$ (eqns. (7.3), (7.7), (7.12)) ;
2 Kalman filter prediction of $\boldsymbol{\mu}_{t|t-1}^{(i)}$, $\boldsymbol{\Sigma}_{t|t-1}^{(i)}$ (eqns. (6.21a) and (6.21b)).;
3 Kalman filter estimation of $\boldsymbol{\mu}_{\mathbf{b},t}^{(i)}$ and $\boldsymbol{\Sigma}_{\mathbf{b},t}^{(i)}$ (eqn. (6.22));
4 Kalman filter correction of $\boldsymbol{\mu}_{t|t}^{(i)}$, $\boldsymbol{\Sigma}_{t|t}^{(i)}$ (eqns. (6.21c) and (6.21d));
5 Evaluation of weights $w_t^{(i)}$ (eqns. (6.38), (6.32));
  end
6 Normalization of importance weights;
7 Resampling;
8 Computation of particle average:

$$\widehat{\mathbf{x}}_t = \sum_{i \in \mathcal{N}} \hat{\mu}_{t|t}^{(i)} \qquad \widehat{\theta}_t = \sum_{i \in \mathcal{N}} \theta_{0:t}^{(i)}, \qquad \widehat{\mathbf{b}} = \sum_{i \in \mathcal{N}} \mu_{\mathbf{b},t}^{(i)}.$$

end

Algorithm 7.1: RBPF using the TVAR model and prior importance sampling of the time-varying source model parameters and noise variance terms.

and eqns. (7.7) and (6.5), such that the source parameters, $\mathbf{a}_t$, the process noise log-variance, ϕ_{v_t}, and the measurement noise log-variance, $\boldsymbol{\phi}_{\mathbf{w}_t}$ are given by

$$p(\mathbf{a}_t \mid \mathbf{a}_{t-1}) = \mathcal{N}\left(\mathbf{a}_t \,\middle|\, \mathbf{a}_{t-1}, \boldsymbol{\Sigma}_{\mathbf{a}_t}\right) \tag{7.3}$$

$$p\left(\phi_{v_t} \mid \phi_{v_{t-1}}\right) = \mathcal{N}\left(\phi_{v_t} \,\middle|\, \phi_{v_{t-1}}, \sigma^2_{\phi_{v_t}}\right) \tag{7.7}$$

$$p\left(\boldsymbol{\phi}_{\mathbf{w}_t} \,\middle|\, \boldsymbol{\phi}_{\mathbf{w}_{t-1}}\right) = \mathcal{N}\left(\boldsymbol{\phi}_{\mathbf{w}_t} \,\middle|\, \boldsymbol{\phi}_{\mathbf{w}_{t-1}}, \sigma^2_{\boldsymbol{\phi}_{\mathbf{w}_t}}\right). \tag{7.12}$$

The RBPF using the TVAR model and based on prior importance sampling of θ_t is therefore summarised in Alg. 7.1, where grey steps are identical to those in the RBPF in Alg. 6.1.

7.3 Demonstration of importance sampling, weighting, and resampling

A 1000 sample long second-order TVAR signal is generated according using the dynamic TVAR parameter model. The Markov variance on the parameters is $\boldsymbol{\Sigma}_{\mathbf{a}_t} = 5 \cdot 10^{-3}\mathbf{I}_Q$ according to the observations in sect. §4.5. Likewise, to enforce smooth variation of the signal, the Markov parameter on the process variance is also $\sigma^2_{\phi_{v_t}} = 5 \times 10^{-3}$.

A single receiving sensor is assumed, i.e., $M = 1$. The source signal is distorted by WGN with variance varying according to a smooth and slowly varying random walk,

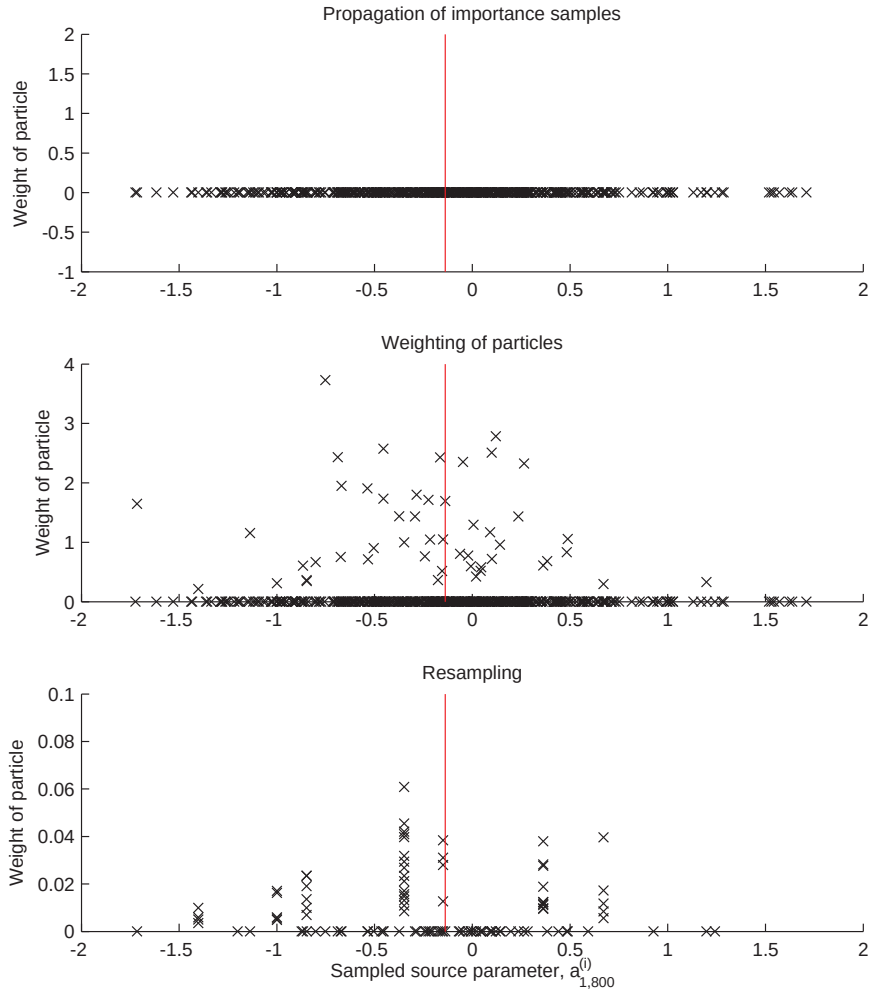


Figure 7.1: *Demonstration of SIR filtering using* 1000 *particles of the source parameter,* $a_{1,t}$*, at* $t = 800$ *of the synthetic data of order* $Q = 2$*, distorted by WGN of SNR* 35*dB and filtered by the AR(8) model of the gramophone horn response. Red lines indicate actual parameter value.*

where $\sigma_{w_t} = 5 \cdot 10^{-3}$. The noise sequence is adjusted to a signal-to-noise ratio (SNR) of 35dB, i.e., a low level of noise to demonstrate the performance of dereverberation rather than denoising. The noisy signal, i.e., the sum of the noise and source signal, is filtered through the AR(8) model of the gramophone horn response.

The RBPF is executed for $N = 1000$ samples assuming the same Markov parameters as for the data generation. At sample $t = 800$, the following stages of the RBPF are plotted in Fig. 7.1:

Importance samples drawn from the prior importance function. As the source model is an AR(2) process, the first source parameter is bounded between $|a_{1,t}| \leq 2$ for stable parameters (recall Fig. 3.15 on page 61 for the admissible region of

second-order autoregressive (AR) processes, following [146]). The majority of importance samples is drawn close to the underlying value of $a_{1,800} = -0.1388$ used for source generation as shown by the histogram of the importance samples in Fig. 7.2. This due to the propagation of accurate particles with time. I.e., assuming that at time $t-1$, the posterior pdf of the source parameters is accurately represented by the particle cloud, then at time t, the importance samples are drawn from dependent on the previous particles. Assuming non-abrupt variation with time of the source parameters, the knowledge inferred from the estimates at $t-1$ is sufficient for a rough estimate at t.

Note that spikes can occur in the histogram, e.g., as found at $a_{1,t} \approx \pm 1$ in Fig. 7.2. These are due to random fluctuations in the particle set and could, for example, be diminished by increasing the bin size of the histogram.

Weighted samples: Particles centred around the true value $a_{1,t} = -0.1388$ are assigned high weights, whereas particles further away, e.g., around 1.5 are of low weight. The point-mass distribution of the particles in this plot approximates the posterior pdf of the source parameter. Its overall envelope resembles that of a Gaussian.

Resampled particles: Particles with low weight are replaced by particles with high weight. Several peaks associated with weights in the point-mass distribution are visible, one of which is located at the actual source parameter location.

This experiment therefore reiterates the concept of source parameter sampling within the RBPF and demonstrated that accurate source model parameters can be evaluated within the proposed framework.

The remainder of this chapter examines the RBPF based on synthetic and speech data distorted by the gramophone horn response. sect. §7.4.1 investigates the performance of the RBPF for synthetically generated source signals. In order to evaluate the baseband performance, a comparison of the estimates of the anechoic source signal, channel and source model parameters, and noise variance terms with their actual, underlying values is necessary. As only the source signal but not the parameters and variance terms are known as *a priori* for speech signals, *synthetic* data generated from the dynamic TVAR parameter model in sect. §7.2 is used instead. Sect. §7.4.2 extends the results to speech signals, demonstrating the accuracy of the RBPF for blind speech dereverberation and the suitability of the TVAR model for speech modelling. Sect. §7.4.3 investigates the performance for different phoneme types and sect. §7.4.4 shows that the dereverberation performance can be improved by using multiple sensors. Conclusions are drawn in sect. §10.5.

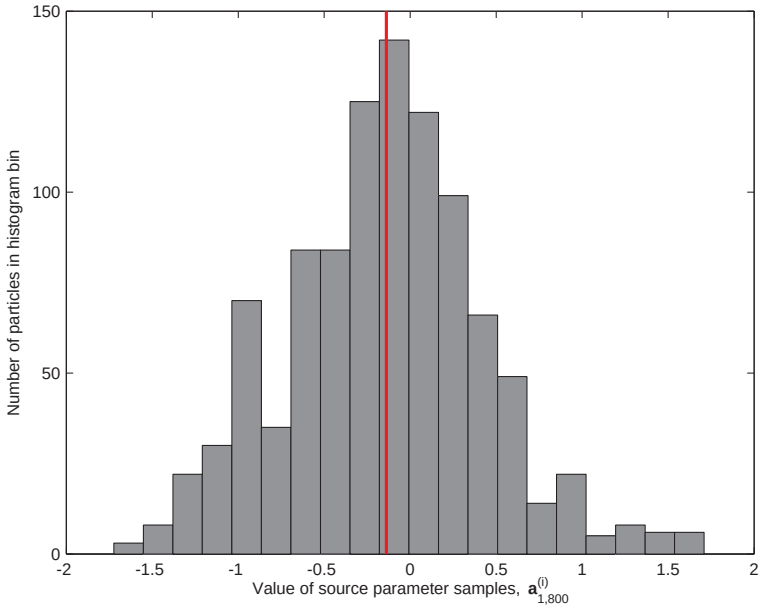


Figure 7.2: *Histogram using* 20 *bins of the importance samples of* $a_{1,t}$ *at* $t = 800$ *for a AR(2) process using* 1000 *particles. Red line indicates actual parameter value.*

7.4 Experimental results

Ideally, the RBPF should be tested based on data measured in reverberant rooms. However, several issues arise with realistic RIRs:

1. As discussed in sect. §4.4.2 on page 73 ff., the model order of the RIR is crucial for accurate modelling of the response. However, in practice, knowledge of the required channel model order is unavailable.
2. The model order of an all-pole filter approximation the RTF accurately is generally unknown and, depending on the sampling frequency, can lie between $100 \leq P \leq 1000$ for *simulated* RIRs using the image-source method (ISM) (see sect. §4.3). The computational burden invoked is therefore prohibitive for computationally efficient dereverberation.

The above-mentioned issues can both be resolved by utilising a multirate extension of the RBPF as proposed in Chap. 10. However, this chapter focuses on the fullband application of the RBPF and therefore an alternative channel to realistic or simulated RTFs is necessary.

Instead of using measured or simulated RTFs, the following results are based on the gramophone horn response discussed in sect. §4.5 on page 74. This response was previously used in the blind dereverberation approach in, e.g., [45], and is therefore considered a viable simplistic reverberant environment for the experiments conducted in the following.

In order to evaluate the improvement of the improvement of the estimated signal over the reverberant signal, the following distortion measures are evaluated:

Segmental signal-to-reverberant component ratio (SRR), measuring the ratio of the signal energy compared to the error energy of the observed or estimated signal over F blocks of length L, i.e., [123]

$$\mathrm{SRR}_{\mathrm{dB}} \triangleq \frac{1}{M} \sum_{f \in \mathcal{F}} 10 \log_{10} \left\{ \frac{\sum_{\ell=Lf}^{Lf+L-1} x_\ell^2}{\sum_{\ell=Lf}^{Lf+L-1} (x_\ell - \chi_\ell)^2} \right\}. \tag{7.13}$$

The segment SRR is therefore the geometric mean of the SRRs across all frames of a speech signal. $\chi_t = y_t$ when evaluating the SRR of the observed signal and χ_t is the estimated source signal sample at time t when evaluated the SRR of the estimated signal. The frame length is typically chosen as $15 - 20$ms for speech signals.

Log-spectral distortion (LSD), measuring the distance between the spectra of the source signal with the observed/estimated signal

$$\mathrm{LSD}_{\mathrm{dB}} \triangleq \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} \left[10 \log_{10} \left(\frac{P(\omega)}{\hat{P}(\omega)} \right)^2 d\omega \right]}, \tag{7.14}$$

where $P(\omega)$ is the spectrum of the source signal and $\hat{P}(\omega)$ is the spectrum of the observed signal or estimated signal.

Bark distortion measure (BSD), a perceptually motivated measure [207] that considers frequency scale warping and critical band integration in the cochlea, changes in sensitivity of the ear as the frequency varies, and considers that loudness is perceived in a non-linear relation to signal intensity [123, 207].

7.4.1 Synthetically generated data according to the source model

A 5000 sample long TVAR signal (in the following referred to as the "source signal") is generated according to the dynamic TVAR parameter model in eqn. (3.19). As speech signals are often modelled using approximately 15 AR coefficients [208, 209], the model order of the synthetic signal is chosen as $Q = 15$. The Markov variance on the parameters is $\boldsymbol{\Sigma}_{\mathbf{a}_t} = 5 \cdot 10^{-3} \mathbf{I}_Q$ according to the observations in sect. §4.5. Likewise, to enforce smooth variation of the signal, the Markov parameter on the process variance is also $\sigma^2_{\phi_{v_t}} = 5 \times 10^{-3}$.

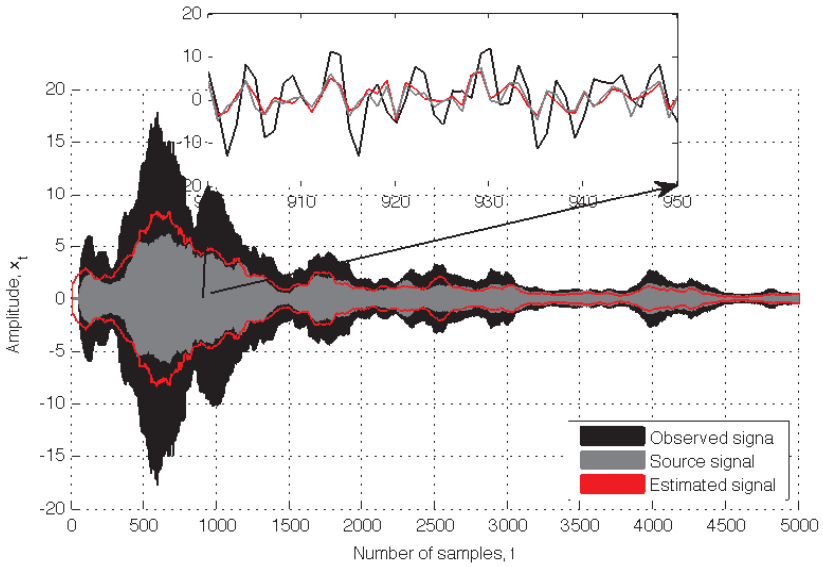


Figure 7.3: *Comparison of the envelopes of the observed, source, and estimated signal.*

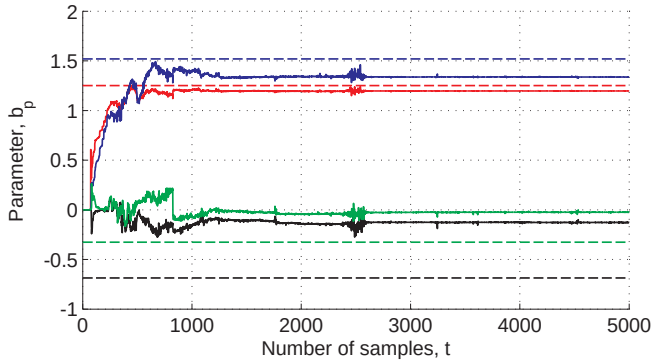
A single receiving sensor is assumed, i.e., $M = 1$. The source signal is distorted by WGN with variance varying according to a smooth and slowly varying random walk, where $\sigma_{w_t} = 5 \cdot 10^{-3}$. The noise sequence is adjusted to a SNR of 35dB. The noisy signal, i.e., the sum of the noise and source signal, is filtered with the $P = 72$-nd order model of the the gramophone horn response in sect. §4.5.

The RBPF is executed using the dynamic TVAR parameter model for $N = 500$ particles. The source and channel order, Q and P, as well as the Markov variance terms $\mathbf{\Sigma}_{\mathbf{a}_t}$ and $\sigma^2_{\phi_{v_t}}$ are assumed known.

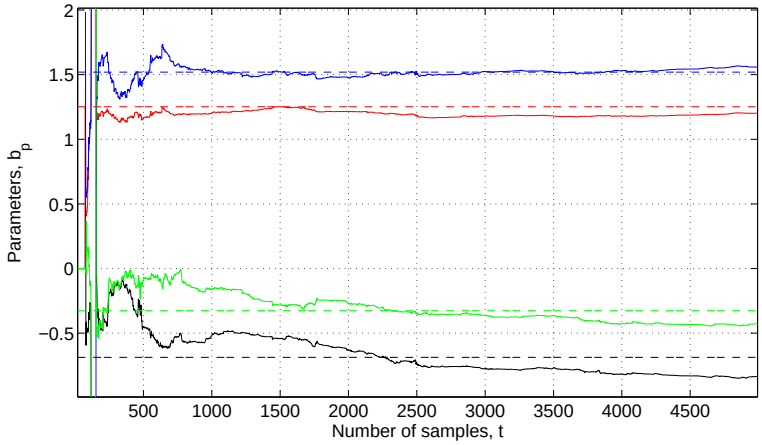
The resulting segmental SRR, LSD and BSD for the RBPF estimate and the reverberant observed signal are summarised in Table 7.1. An improvement of $\Delta\,\mathrm{SRR}_{\mathrm{dB}} = 8.99$dB from a segmental SRR of -4.89dB for the observed signal to 4.10dB of the RBPF estimate is achieved. Furthermore, the LSD is from 1.35dB of the observed signal to 0.56dB of the estimate signal, i.e., an improvement of 0.79dB or 58% relative to the observed signal LSD.

	Segmental SRR	LSD	BSD
Observed signal	-4.89	1.35	0.17
RBPF estimate	4.10	0.56	0.01
Improvement	8.99	0.79	0.16

Table 7.1: *Distortion measures for synthetic data comparing the RBPF estimate and observed signal distortion.*



(a) *Comparison of actual (dashed) and estimated (solid) channel parameters for* b_1 *(red),* b_3 *(blue),* b_6 *(black) and* b_8 *(green).*



(b) *Experiment in Fig. 7.4a for* 10 *Monte Carlo runs.*

Figure 7.4: *Convergence of the channel parameter and pole estimates with the actual channel.*

The improvement in the signal quality is visible in the time-domain signal. The envelopes of the observed, source, and estimated signals are shown in Fig. 7.3. The zoomed-in section between $900 - 950$ samples highlights that the estimated signal accurately approximates the dynamic behaviour of the source signal. The estimated signal requires approximately 1000 samples to converge towards the source signal. As particle filters are sequential estimators, the same estimation performance should be evident for the whole cycle of samples and so-called 'burn-in' periods, observed in Markov chain Monte Carlo approaches, should not occur. However, as shown in Fig. 7.4a, the channel parameter estimates require approximately 1500 samples to converge towards a steady state. The corresponding 95^{th} convidence intervals for b_3 and b_8 are shown in Fig. 7.4b. Fig. 7.4b shows the convergence of the channel parameters for a repetition of the same experiment over 10 Monte Carlo runs. Fig. 7.4b therefore demonstrates that the parameters converge to the same values for multiple runs of the

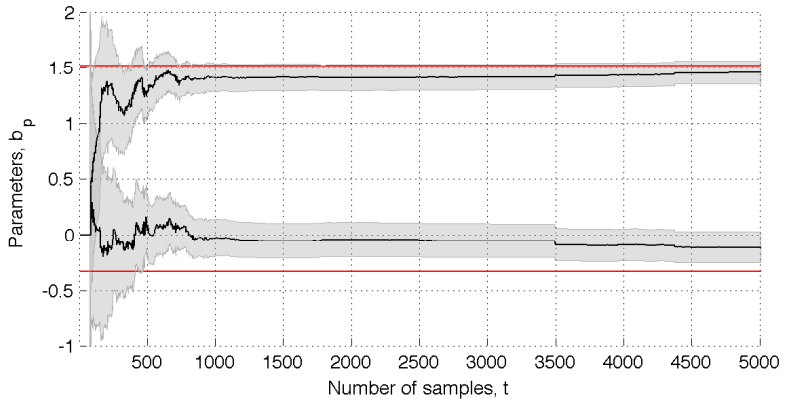
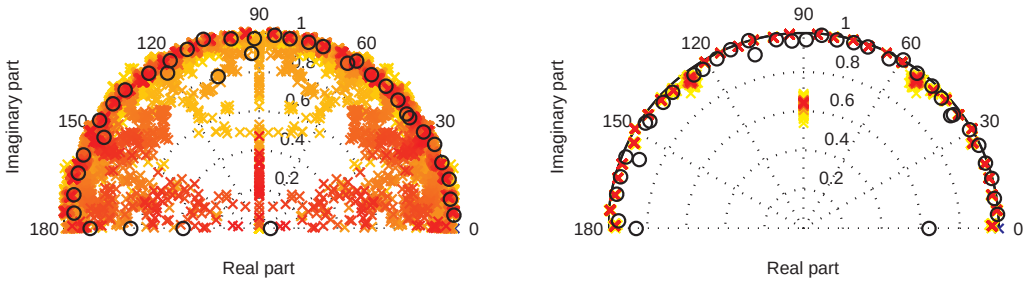


Figure 7.5: *Confidence interval (grey area) of estimated parameters* b_3 *(top) and* b_8 *(bottom) for* 1 *Monte Carlo run vs. actual parameters (red).*

algorithm.

As the source signal estimation is dependent on the channel estimates, inaccurate channel estimates can cause inaccurate source signal estimates. Thus, the initial 'burn-in' period of the source signal is due to the convergence time required for the channel estimates. This claim can be confirmed by executing the RBPF for the same data and setup assuming that the channel parameters are known.

Similar to the parameters, the channel poles converge towards the pole locations of the horn response within approximately 1500 samples. Fig. 7.6a shows that the initial estimated poles between $t = 1, \ldots, 625$ are scattered over the unit circle. Nonetheless, after the channel estimate has converged, the poles approximate those of the gramophone horn model accurately, as shown by the trajectory of poles between $t = 1, 4900, \ldots, 5000$ in Fig. 7.6b. It should be noted at this point that the poles are



(a) *Initial trajectory of poles between* $t = 1$ *(red) and* $t = 625$ *(yellow) compared to the actual poles (black circles).*

(b) *Trajectory of poles between* $t = 4900$ *(yellow) to* $t = 5000$ *(red) samples.*

Figure 7.6: *Convergence of the channel parameter and pole estimates with the actual channel.*

	Segmental SRR	LSD
Observed signal	-6.59	1.67
LP residual technique [79]	2.97	1.59
RBPF estimate	2.97	1.35
Improvement of RBPF over observed signal	9.56	0.31
Improvement of RBPF over [79]	0	0.24

Table 7.2: *Distortion measures for speech data comparing the RBPF estimate and observed signal distortion.*

located very closely to each other. Therefore, the convergence of some parameters towards a false solution could be due to identifiability issues between the closely positioned poles.

7.4.2 Speech data

The experiment is repeated for the 2.62s long sentence "In the long run, it pays to buy quality clothing." uttered by a male American and downsampled to 4kHz. Again, the signal is distorted by WGN of SNR 35dB. Due to the low sampling frequency, the high frequency components of the 72-nd order channel model of the 11.025kHz gramophone horn response would not taken into account. Instead, the 8-th order model is used instead to filter the noisy speech signal.

The RBPF is run for 1000 particles, again using $Q = 15$ source and $P = 8$ channel parameters. The resulting distortion measures are summarised in Table 7.2. The RBPF achieves an improvement of $\Delta\,\mathrm{SRR}_{\mathrm{dB}} = 9.56$dB compared to the observed signal SRR of -6.59dB to an estimated SRR of 2.97dB. The LSD is improved by 0.31dB from 1.67dB in the observed signal to 1.35dB in the estimated signal. This improvement can be particularly well demonstrated on the actual speech signals. Audio samples of the anechoic speech, observed, and estimated signal can be found on the attached compact disc (CD) in the folder '*Chapter 5-6 - TVAR and PFS model*'. The observed signal exhibits metallic distant sound, whereas the estimated signal reduces the metallic sound significantly.

For a comparison to existing blind dereverberation techniques, the same signal is processed using the linear prediction (LP) residual technique by Yegnanarayana and Satyanarayana Murthy [79] as discussed in sect. §2.5 on page 23.[1] A LPC order of $P = 10$ and a frame length of 20ms is assumed. As summarised in Table 7.2, the LP residual technique results in a segmental SRR of 2.97dB and a LSD of 1.59. An audio sample of the estimated signal can be found on the attached CD in the folder '*Chapter*

[1]The author would like to thank Dr Nikolay D. Gaubitch for the implementation of the algorithm.

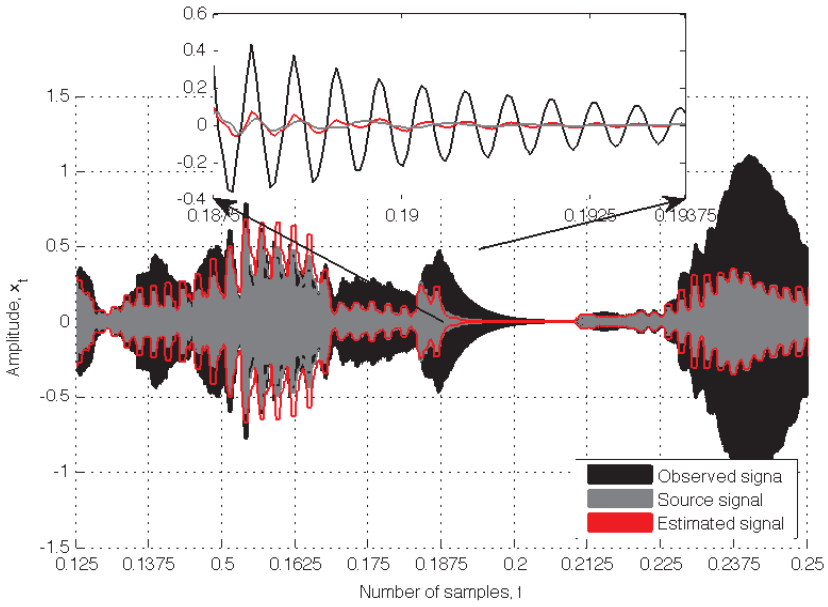


Figure 7.7: *Comparison of the anechoic, reverberant, and estimated speech signal, "In the long run, it pays to buy quality clothing." at* $f_s = 4kHz$*, between* $0.2 - 0.5s$*.*

5-6 - TVAR and PFS model', demonstrating that the improved signal still retains significant metallic sound effects from the horn response.

Hence, the RBPF performs as well as the LP technique in [79] in terms of the segmental SRR. Furthermore, the RBPF achieves improvement of 0.24dB compared to the technique in [79]. The provided audio samples demonstrate improved quality of speech of the RBPF estimate as compared to the signal processed by the LP residual technique. The significant improvement in audio quality of the RBPF over [79] despite equal segmental SRR demonstrates that the ability to predict *subjective* speech quality using the segmental SNR is limited due to the lack of modelling of the auditory system. This is partially due to the fact that the signal energy during intervals of silence (e.g., breathing pauses or short breaks between words) is small, such that the segmental SNR in intervals of silence can take large negative values, thus biasing the overall measure. In order to circumvent this effect, silent frames should either be excluded from the signal, or the segmental SNR should be floored to a small value, e.g., by limiting SNR_{seg} in a range of $[-10dB, 35dB]$ in order to avoid voice activity detection [210].

The improvement of the estimated signal over the reverberant signal and the close approximation of the underlying anechoic speech signal are reflected by the comparison of the time-domain signals. A zoomed in version plotting the corresponding

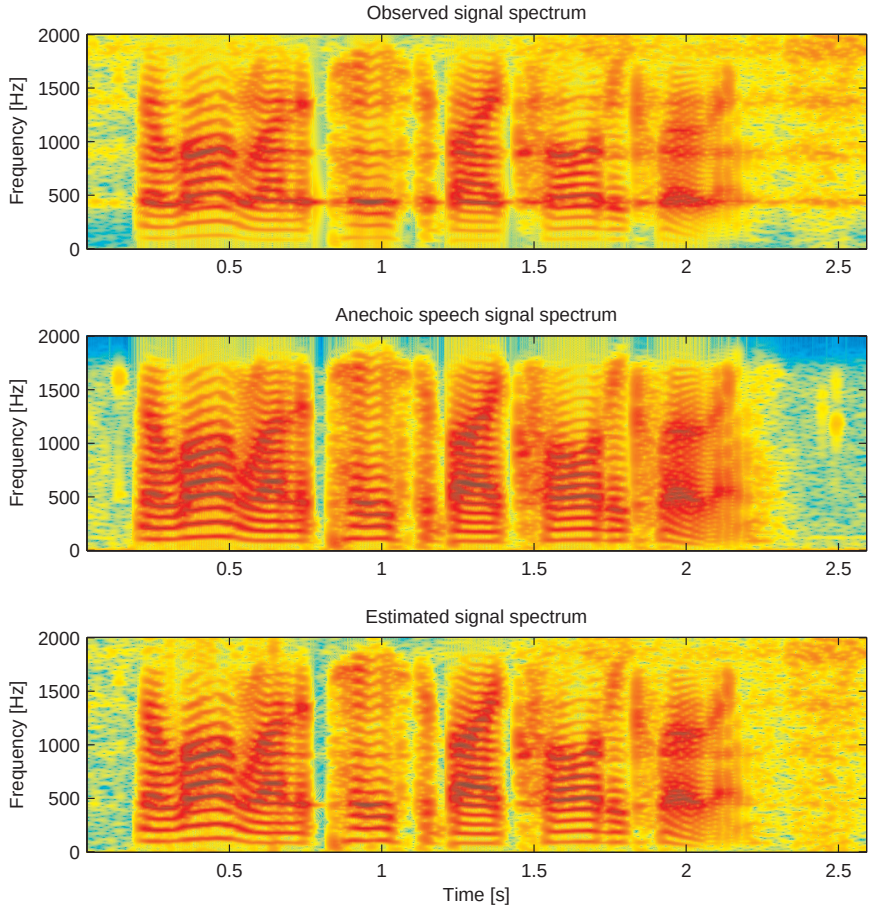


Figure 7.8: *Spectrograms of the anechoic, reverberant, and estimated speech signal, "In the long run, it pays to buy quality clothing." at* $f_s = 4kHz$*. Color scale from dark red (high energy) to dark blue (low energy) indicates the dynamic range of the signal between* $1dB$ *and* $-7dB$*.*

signals between 0.125 and 0.25s is shown in Fig. 7.7. As illustrated in more detail between $0.1875 - 0.1937$s, the anechoic is approximated accurately by the estimated signal. Even at abrupt changes in the speech variation (see, e.g., at 0.225s), the speech model adapts quickly to the changes in variability, leading to accurate estimates.

The spectrograms of the anechoic, reverberant and estimated signals are plotted in Fig. 7.8. Comparing the observed and anechoic source signal (top and centre graph), reverberation and noise mainly remove high energy components from the baseband between $0 - 500$Hz and introduce high energy in the subband between $1.75 - 2$kHz, where the anechoic signal consists mainly of low energy components. Furthermore, energy is smeared in the observed spectrogram into the anechoic low-energy time-segment between $2.25 - 2.6$s. The estimated signal partially reconstructs the high-energy components in the baseband between $0 - 500$Hz and reduces the high energy

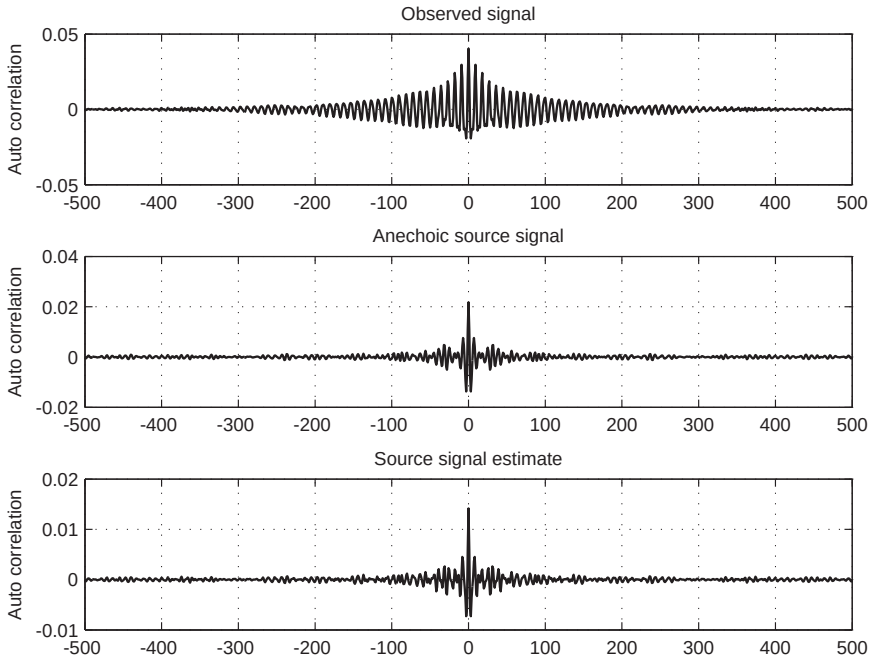


Figure 7.9: *Autocorrelation functions of the observed signal (top), anechoic speech signal (centre), an estimated signal (bottom).*

components in the time-segment between $2.25 - 2.6$s.

The smearing of high energy components into the last time-segment is illuminated upon by Fig. 7.9, illustrating the autocorrelation of the three signals, i.e., the similarity between samples as a function of the lag between them. It can be seen that the observed signal is correlated up to lags of 250 samples, whereas the source signal shows autocorrelations of up to approximately 75 samples. The effect of increased autocorrelation in the observed signal is due to the all-pole filter operation of the channel, relating the current samples to a linear combination of past sample trajectories. This also explains the smearing of high energy components into the low-energy time segment in the spectrogram of the observed signal. As can be seen from the autocorrelation function of the estimated signal, the short-lagged autocorrelation of the anechoic source signal is restored to a large extent. Hence, the high-energy components in the last time-segment of the estimated spectrogram are reduced as compared to the observed signal.

7.4.3 Investigation of performance for different phoneme types

In order to investigate the suitability of the model for different phoneme types in speech signals, four types of speech sounds are extracted from a database of 10 sentences uttered by a female American speaker from the Texas Instruments, Inc., and

Phoneme type		Segmental SRR	LSD
Fricatives	Observed signal	−3.11	1.19
	RBPF estimate	2.14	0.82
	Improvement	**5.25**	**0.37**
Stop consonants	Observed signal	−2.65	1.261
	RBPF estimate	4.16	1.268
	Improvement	**6.81**	**−0.007**
Vowels	Observed signal	−2.70	1.17
	RBPF estimate	−0.88	1.79
	Improvement	**1.82**	**−0.62**
Semivowels	Observed signal	−4.19	1.28
	RBPF estimate	0.42	1.74
	Improvement	**4.61**	**−0.46**

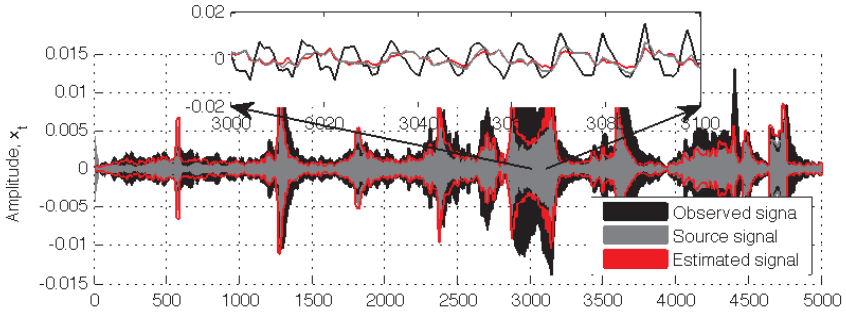
Table 7.3: *Distortion measures for speech data comparing the RBPF estimate and observed signal distortion.*

Massachusetts Institute of Technology (TIMIT) database at $f_s = 16$kHz, i.e.,

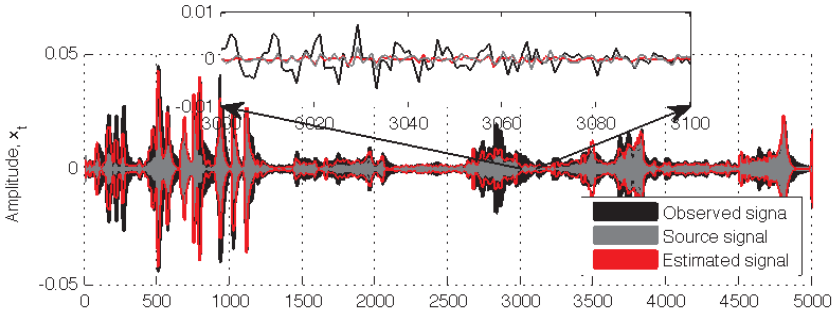
1. vowels (e.g., /iy/, /ae/),
2. semivowels (e.g. /r/, /l/),
3. fricatives (e.g., /sh/, /z/), and
4. stop consonants (e.g., /b/, /d/).

The phoneme types are concatenated into four separate synthetic speech sequences, each of which contains only one sound type. The signals are downsampled to $f_s = 4$kHz and distorted by WGN of SNR 35dB and convolved with the 8-th order gramophone horn model.

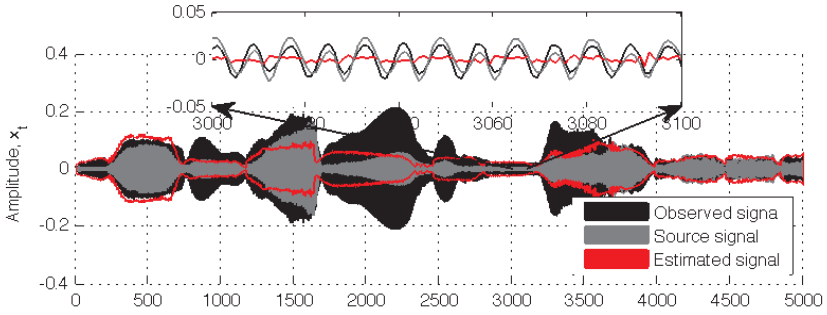
The four sound type sequences are processed using the RBPF using 1000 particles. Speech is typically modelled using 15 AR parameters. Hence, the RBPF assumes $Q = 15$ and $P = 8$. The resulting SRR and LSD values are summarised in Table 7.3. The dynamic TVAR parameter model is particularly apt at modelling fricatives and stop consonants with SRR improvements of 5.25dB and 6.81dB respectively. However, issues are encountered with modelling vowels and semivowels. For both types, the estimates cause LSDs degradations. For vowels, the SRR improvement is as little as 1.82dB. The modelling accuracy of fricatives and stop consonants and inaccuracy for vowels and semivowels can be illustrated well by means of the time-domain signal plots in Fig. 7.10. The envelopes of the anechoic signals for the fricative sounds and stop consonants are accurately captured by their corresponding estimates. However, the estimated envelopes for the vowels and semivowels seem to rather attenuate the overall observed signal than model that of the anechoic source signal.



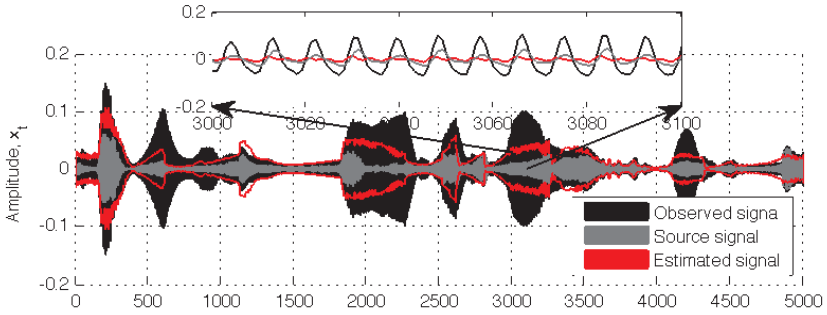
(a) *Fricative sounds*



(b) *Stop consonants*



(c) *Vowels*



(d) *Semivowels*

Figure 7.10: *Comparison of the anechoic source, observed, and estimated signals for four speech sequences containing either fricatives, stop consonants, vowels, or semivowels.*

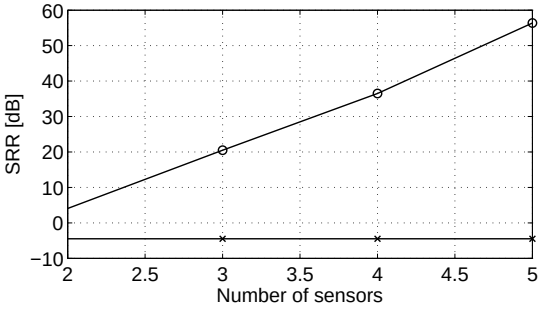


Figure 7.11: *Improvement in SRR by using multiple sensors*

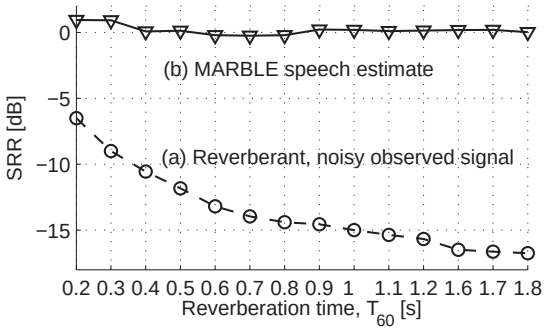
7.4.4 Improving estimates by using multiple sensors

The estimates can be improved upon by using multiple sensors, $M > 1$, in order to exploit spatial diversity. As the source-sensor positions are different for each sensor, $m \in \mathcal{M}$, the RIR between each sensor and the source is unique. Hence, the channel parameters characterising the RIR differ between sensors.

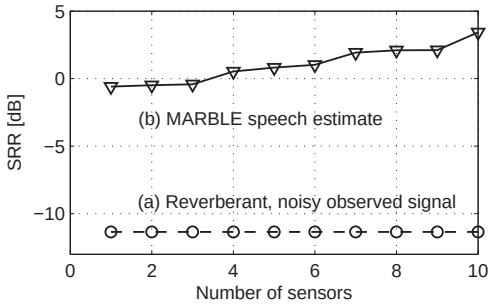
To illustrate this, consider perturbing the pole positions of the 8-th order all-pole model of the gramophone horn to generate $M = 5$ distinct variants of the channel response. The channel between the first sensor, $m = 1$, and the source is assumed to be the original 8-th order response. For the remaining four channels, the radii of the poles are randomly perturbed by a WGN with variance 0.0005. The phases are perturbed with a variance of 0.005. Corresponding poles lying outside of the unit circle are redrawn until a stable channel model is generated. A synthetic source signal is generated as described in sect. §7.4.1 for $Q = 15$ source parameters and is filtered with each of the five channel responses. Four experiments of the RBPF are then executed using 500 particles for $M = 2, \ldots, 5$ sensors.

The improvement in the segmental SRR for the experiments is shown in Fig. 7.11. The results indicate that the performance of the RBPF can be significantly increased by using multiple microphones.

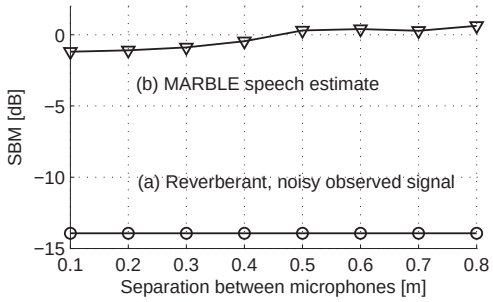
The experiment based on perturbations of the gramophone horn response, is, however, a rather artificial example for multi-sensor processing. Conclusive results about the exact improvement of the enhancement performance should therefore be made using realistic impulse responses instead. However, realistic RIRs are typically characterised by $600 - 1000$ poles for fullband signals of sampling frequencies between $8 - 16$kHz and depending on the reverberation time and dimensions of the room. Estimation of 1000 resonant poles close to the unit circle not only leads to identifiability issues between poles but also induces a significant computational burden as will be



(a) *Increasing reverberation time,* T_{60}.



(b) *SRR improvement with increasing number of sensors*



(c) *SRR improvement with increasing separation between sensors*

Figure 7.12: *Indicative results of the subband RBPF performance for increasing reverberation time, number of sensors and separation between sensors.*

discussed in Chap. 10. The same chapter thus proposes that *subband methods* should be used for blind speech dereverberation of realistic RIRs, e.g., using responses generated by the ISM in sect. §4.3.

Therefore, for indicative results for realistic RIRs, a shoe-box room of dimensions $2.78 \times 4.68 \times 3.2$m (width $\times$ depth $\times$ height) is simulated using the ISM at a subband of $f_s = 500$Hz for $T_{60} = 0.2$s using a setup similar to Fig. 9.4 on page 178: one sensor is initially placed on the East side of the room on table height, where $\mathbf{r}_o = \begin{bmatrix} x_o^T & y_o^T & z_o^T \end{bmatrix}^T$ where $x_o = 3.265$m, $y_o = 1.17$m and $z_o = 1.3$m. The source is located on the West side of the room at an average adult's heigh, i.e., $\mathbf{r}_s = \begin{bmatrix} x_s & y_s & z_s \end{bmatrix}^T$ where $x_s = 1.54$m, $y_s = 1.98$m and $z_0 = 1.7$m.

The sentence "She had your dark suit in greasy wash water all year" from the TIMIT database uttered by a male American speaker is down-sampled to 500Hz and distorted by WGN varying according to a random walk and of SNR 25dB. The RBPF is executed for 50 particles over 10 Monte Carlo runs. The experiment is repeated for T_{60} times in increments of 0.2s until $T_{60} = 1.8$s. The segmental SRR is computed for

each set of results and is plotted in Fig. 7.12a. Whilst the SRR of the observed signal decays from approximately -6dB for $T_{60} = 0.2$s to -19dB for $T_{60} = 1.8$s, the SRR of the estimated signal only decays by about 1dB between reverberation times between $0.2 - 1.8$s. The marginal degradation of the RBPF estimate SRR compared to the rapid degradation in the observed signal SRR therefore suggests that the RBPF is comparatively robust against increases in reverberation time.

Next, the experiment is repeated for $T_{60} = 0.45$s for the number of sensors increasing over $M = 1, \ldots, 10$. Again, the resulting SRR of the observed signal and RBPF are plotted in Fig. 7.12b. The SRR of the signal increases from 0dB for a single sensor to 5dB for ten sensors. Multi-sensor processing can therefore lead to significant SRR improvement. The performance of multi-sensor processing can be further improved by increasing the inter-sensor distance in order to exploit spatial diversity. As indicated in Fig. 7.12c, the performance for $M = 2$ sensors can be improved by approximately 2dB by increasing the sensor separation by 0.6m.

7.5 Discussion

This chapter introduced a dynamic TVAR speech parameter model for the RBPF, where the source parameters are assumed to vary according to a smooth random walk. Experimental results were presented on synthetic and speech data convolved with a gramophone horn response. Results demonstrated segmental SRR improvements of a reverberant signal 9.3dB and an LSD improvement of 0.31dB. Audio samples demonstrated the improvement in signal quality of the estimated signal over the observed signal.

A survey on individual speech sounds showed that the dynamic TVAR parameter model is particularly apt at modelling fricatives and stop consonants. However, the model struggles with vowels and semivowels in particular. Therefore, the TVAR parameter model accurately models *unvoiced* speech. Enhancement can be improved by accurately modelling the speech production mechanism for *voiced* speech sounds. Chap. 8 therefore proposes a novel speech model based on parallel formant synthesis.

Chapter 8

Articulatory-based speech model using parallel formant synthesis

8.1 Introduction

The dynamic time-varying AR (TVAR) source parameter model in Chap. 7 facilitates accurate blind speech dereverberation for fricatives and stop consonants, i.e., for unvoiced speech. However, no prior information besides the random walk on the TVAR parameters is assumed that would otherwise bias the model to follow a specified variation, such as harmonic or plosive behaviour. Therefore, the TVAR parameter model can be improved upon by exploiting prior knowledge about the human speech production mechanism in order to improve blind dereverberation for quasi-periodic, voiced speech segments.

As discussed in sect. §3.4.3, parallel formant synthesizers (PFSs) are popular speech models in the speech *synthesis* community for modelling the formants of the vocal tract by means of a concatenation of several resonator circuits. Formants are defined as the spectral peaks in the speech spectrum [211], corresponding to the resonances in the human vocal tract. Therefore, this chapter proposes to parameterise the TVAR source model from a PFS perspective. As resonator circuits can be described using second-order autoregressive (AR) models, the PFS can be written in the state space form required by the Rao-Blackwellized particle filter (RBPF) in eqn. (6.9) on page 115. In the speech synthesis community, the AR process is driven by the resonant frequency and bandwidth which are set manually using an amplitude control driving the resonator circuit. In speech estimation, knowledge of the resonant frequency and bandwidth are not available. Thus, an appropriate parameter model in terms of the resonant frequency and bandwidth is required.

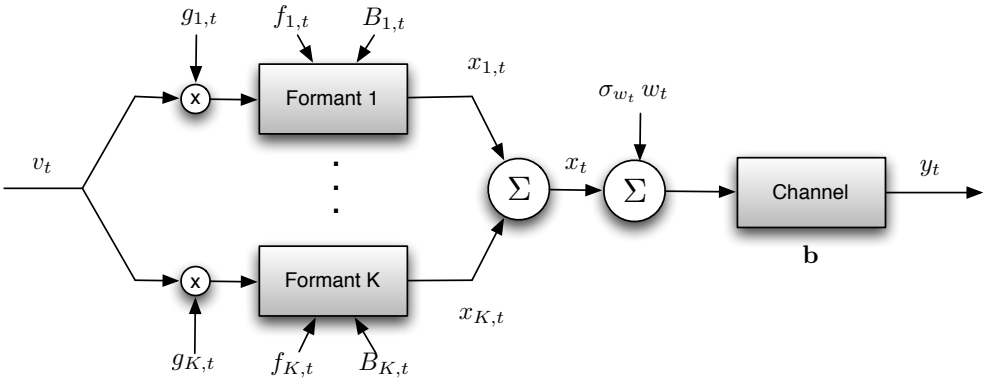


Figure 8.1: *Parallel formant synthesiser.*

This chapter therefore investigates the relation between the TVAR parameters and the frequencies and bandwidths of the resonators. Based on these results, a novel parameter model in terms of the partial correlation (PARCOR) coefficients of the resonator circuit is developed, enforcing both stable parameters and resonant frequencies between 0 and π corresponding to valid 3dB bandwidths.

The resulting model is incorporated in the RBPF and results for speech data and different phonemes are presented and compared to the results obtained using the dynamic TVAR parameter model in Chap. 7. Results indicate improved enhancement of the speech signal for vowels, fricatives and consonants using the proposed PFS model over the dynamic TVAR parameter model.

This chapter is thus structured as follows: Sect. §8.2 derives the TVAR signal model of the parallel formant synthesizer. Sect. §8.3 motivates the investigation for a novel PFS parameter model. Sect. §8.4 investigates the relationship between stable TVAR parameters and resonant frequencies corresponding to valid 3dB bandwidths in the magnitude response. Sect. §8.5 derives the proposed novel parameter model based on a random walk on the PARCOR coefficients of the resonator circuit. Sect. §8.6 presents results for speech data and a comparison to the results in Chap. 7. Conclusions are draw in sect. §8.7.

8.2 System model

PFSs simulate the formants of the human vocal tract by means of several resonant circuits connected in parallel as discussed in sects. §3.2.2 and §3.4.3 on page 36 and on page 57 and illustrated in Fig. 8.1. Each resonator signal obeys the second-order TVAR

process in eqn. (3.36) on page 58, i.e.,

$$x_{t,k} = a_{1,t,k}\, x_{t,k-1} + a_{2,t,k}\, x_{t,k-2} + g_{t,k}\, v_{t,k}, \qquad v_t \sim \mathcal{N}(0,\, 1)\,. \tag{8.1}$$

where $k \in \mathcal{K}$ and where K is the number of resonators in the PFS model, $g_{t,k}$ is the gain of the resonator, and $\mathbf{a}_{t,k} = \begin{bmatrix} a_{1,t,k} & a_{2,t,k} \end{bmatrix}^T$ are the TVAR parameters describing the second-order resonator circuit. In speech *synthesis*, each resonator circuit is driven by an amplitude control, defining the resonant frequency and bandwidth, $f_{t,k}$ and $B_{t,k}$ respectively, of the resonator. As, by definition, resonator circuits have resonant poles, i.e., poles close to the unit circle, $f_{t,k}$ and $B_{t,k}$ can be related to the TVAR parameters via the expression

$$\hat{f}_{t,k} \approx \frac{f_s}{2\pi}\phi_{t,k} \qquad \Leftrightarrow \qquad \phi_{t,k} \approx \frac{2\pi}{f_s}\hat{f}_{t,k} \tag{8.2a}$$

$$\hat{b}_{t,k} \approx -\frac{f_{t,k}\ln(r_{t,k})}{\pi} \qquad \Leftrightarrow \qquad r_{t,k} \approx \exp\left\{-\frac{\pi\,\hat{b}_{t,k}}{f_{t,k}}\right\}. \tag{8.2b}$$

where $r_{t,k}$ is the radius $\phi_{t,k}$ is the phase of the complex poles corresponding to the TVAR parameters, which can be related via eqn. (3.45) on page 61, i.e.,

$$a_{1,t,k} = -2r_{t,k}\cos\phi_{t,k} \qquad a_{2,t,k} = r_{t,k}^2 \tag{8.3}$$

It is important to note that eqn. (8.2a) is only valid for poles close to the unit circle, i.e., $r_{t,k} \approx 1$. Eqn. (8.2b) is generally valid for a one-pole system only according to sect. §3.4.3.3 on page 59, as discussed in sect. §3.4.3.2 on page 59. Again, eqn. (8.2b) approximates the bandwidth of resonator circuits for $r_{t,k} \approx 1$ only.

Eqn. (8.1) can be easily written in state-space form by augmenting the last $Q = 2$ resonator signals, i.e.,

$$\underbrace{\begin{bmatrix} x_{t,k} \\ x_{k,t-1} \end{bmatrix}}_{\mathbf{x}_{t,k}} = \underbrace{\begin{bmatrix} a_{1,t,k} & a_{2,t,k} \\ 1 & 0 \end{bmatrix}}_{\mathbf{A}_{t,k}} \underbrace{\begin{bmatrix} x_{k,t-1} \\ x_{k,t-2} \end{bmatrix}}_{\mathbf{x}_{k,t-1}} + \underbrace{\begin{bmatrix} \sigma_{\mathbf{v}_{t,k}} & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{\Sigma}_{\mathbf{v}_{k,t}}} \underbrace{\begin{bmatrix} v_{t,1} \\ v_{t,2} \end{bmatrix}}_{v_{t,k}}, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}_{2\times 1},\, \mathbf{I}_2)\,. \tag{8.4}$$

The PFS source model can be expressed in the form required by eqn. (6.9) on page 115 by augmenting all K resonator output in one state space, i.e.,

$$\mathbf{x}_t = \mathbf{A}_t\mathbf{x}_{t-1} + \mathbf{\Sigma}_{\mathbf{v}_t}\mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}_{2\times 1},\, \mathbf{I}_2)\,, \tag{8.5}$$

where $\{a_{q,t,k} : q \in \mathcal{Q},\, k \in \mathcal{K}\}$ are given by eqn. (8.2), and the states are defined as $\mathbf{x}_t \triangleq$

$\begin{bmatrix} x_{1,t} & x_{1,t-1} & \dots & x_{K,t} & x_{K,t-1} \end{bmatrix}^T$. The source transition model is given by

$$\mathbf{A}_t \triangleq \operatorname{diag}\begin{bmatrix} \mathbf{A}_{1,t} & \dots & \mathbf{A}_{K,t} \end{bmatrix} = \begin{bmatrix} a_{1,t,1} & a_{2,t,1} & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & a_{1,t,K} & a_{2,t,K} \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \tag{8.6}$$

and the process noise matrix is composed of

$$\boldsymbol{\Sigma}_{\mathbf{v}_t} \triangleq \begin{bmatrix} \sigma_{v_{1,t}} & 0 & \cdots & \sigma_{v_{t,k}} & 0 \\ & & \mathbf{0}_{1\times 2K} & & \end{bmatrix}^T. \tag{8.7}$$

The speech signal is constructed by summing over the resonator outputs as expressed by eqn. (3.37) on page 58, i.e.,

$$x_t = \sum_{k\in\mathcal{K}} x_{t,k}. \tag{8.8}$$

The source model of the PFS model in eqn. (8.5) is hence of the form as required for the RBPF in eqn. (6.9) in on page 115.

As before, the set of unknown variables in the system consists of $\boldsymbol{\varphi}_t = \begin{bmatrix} \mathbf{z}_t^T & \theta_t \end{bmatrix}^T$. The analytically tractable variables now contain the resonator outputs of the PFS model and the channel model, i.e., $\mathbf{z}_t = \begin{bmatrix} \mathbf{x}_t^T & \mathbf{b}^T \end{bmatrix}^T$. As the source parameters in eqn. (8.5) are determined by the resonant frequency and bandwidth and the process covariance corresponds to the resonator gain, the time-varying, untractable unknown variables are defined as $\theta_t = \begin{bmatrix} \mathbf{f}_t^T & \mathbf{B}_t^T & \mathbf{g}_t^T \end{bmatrix}^T$, where the vector of resonant frequencies is defined as $\mathbf{f}_t = \begin{bmatrix} f_{1,t} & \dots & f_{K,t} \end{bmatrix}^T$, the bandwidths are given by $\mathbf{B}_t = \begin{bmatrix} B_{1,t} & \dots & B_{K,t} \end{bmatrix}^T$, and the gains are $\mathbf{g}_t = \begin{bmatrix} g_{1,t} & \dots & g_{K,t} \end{bmatrix}^T$. It is therefore desirable to specify the proposal distribution of the RBPF in terms of the resonant parameters, $\mathbf{f}_t$, $\mathbf{B}_t$ and $\mathbf{g}_t$, rather than the TVAR parameters.

8.3 Sampling of the resonant frequency and bandwidth

Sect. §8.2 introduced the speech *signal* model of the PFS model. Similar to the TVAR model in Chap. 7, the dynamic of the modelled speech signal is determined by the model parameters. As the PFS model is parameterised in terms of $f_{t,k}$ and $B_{t,k}$, models on the resonant frequency and bandwidth are required.

In order to investigate the time-varying behaviour of the formant frequencies, the $Q = 12$ AR coefficients of the speech model are extracted from the sentence "She had your dark suit in greasy wash water all year." uttered by a female American speaker at $f_s = 16$kHz with a duration of 3.87s (i.e., $62,055$ samples). Using the twelve extracted TVAR coefficients, the speech sequence is modelled as an AR process of 8000 samples length. In other words, an all-pole filter characterised by the extracted TVAR coefficients is excited by white Gaussian noise (WGN) to generate a synthetic signal. The corresponding magnitude responses varying with time are shown in Fig. 8.2. As the process is of order twelve, $Q/2 = 6$ formants (or spectral peaks) are extracted from the magnitude response using the approach in [212]. The resulting trajectories of resonant frequencies with time over a period of 1000 samples are shown in Fig. 8.3.

From Fig. 8.3 it can be seen that the resonant frequencies vary relatively slowly with time, with a variation of up to 500Hz and spectral magnitudes variations by up to 10dB between $t = 1$ and $t = 800$. The slow variation suggests that the resonant frequencies could be modelled as a random walk. Furthermore, as $g_{t,k}$ in eqn. (3.36) is equivalent to σ_{v_t} in eqn. (3.18) on page 48, the resonant gain can be modelled as a random walk as well according to eqn. (3.13) on page 42. Beierholm and Winther [111] therefore propose to sample the resonant frequencies, bandwidths, and resonant gain from a first-order Markov chain, i.e.,

$$f_{k,t} = f_{k,t-1} + \delta_f r_{f_{k,t}}, \qquad r_{f_{k,t}} \sim \mathcal{N}(0, 1) \tag{8.9a}$$

$$B_{k,t} = B_{k,t-1} + \delta_b r_{B_{k,t}}, \qquad r_{B_{k,t}} \sim \mathcal{N}(0, 1) \tag{8.9b}$$

$$\ln g_{k,t} = \ln g_{k,t-1} + \delta_g r_{g_{k,t}}, \qquad r_{g_{k,t}} \sim \mathcal{N}(0, 1). \tag{8.9c}$$

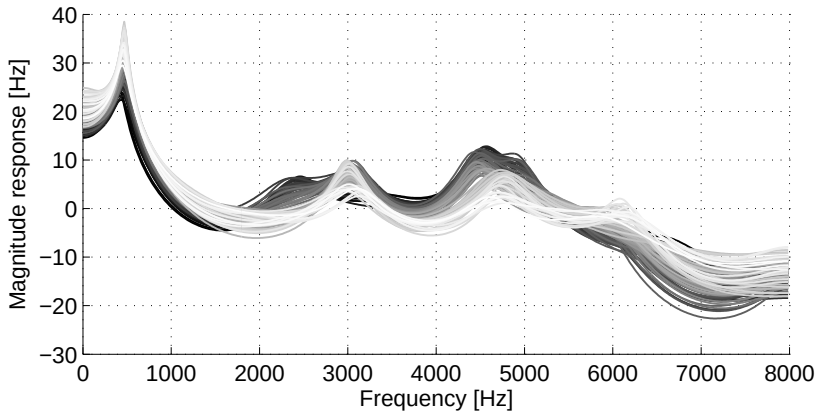


Figure 8.2: *Variation of the spectral magnitude response between* $t = 1$ *(black) and* $t = 8000$ *(white).*

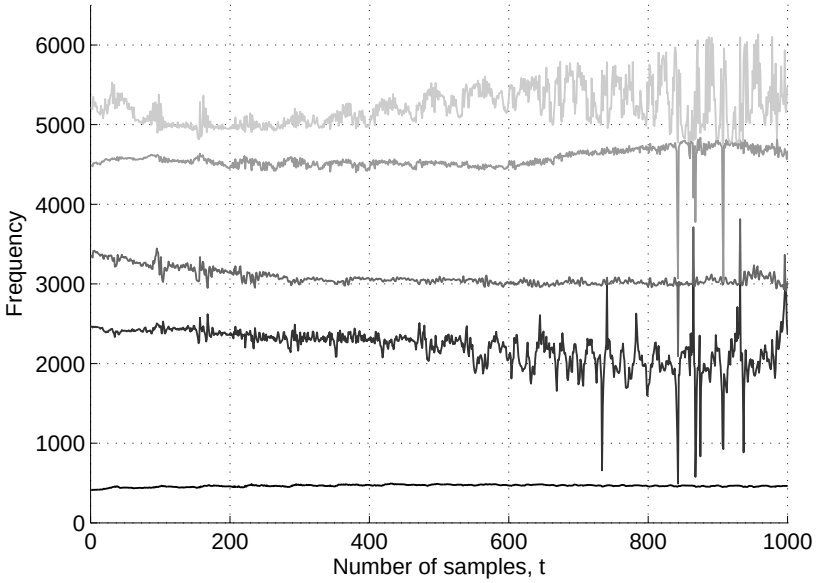


Figure 8.3: *Variation of the first five formant frequencies during the utterance 'iy' by a female speaker at* $f_s = 16kHz$.

where the Markov parameters on the frequency, bandwidth and gain, δ_f, δ_B, and δ_g, are assumed constant and known. Note that the random walk is modelled on the log-gain due to the relation between the variance and the gain in eqn. (3.36).

In order to relate the resonant frequency and bandwidth to the TVAR parameters, Beierholm and Winther [111] utilise the relation in eqn. (8.2) on page 149. However, eqn. (8.2) is only valid for poles close to the unit circle. Whilst this is a valid assumption in speech synthesis, where the resonant frequency and bandwidth are known and necessarily generate resonant poles, this assumption is not necessarily true for speech estimation. In fact, according to eqn. (8.9), the poles can be located anywhere inside and outside of the unit circle. Furthermore, the random walk in eqn. (8.9) is not constrained to generate valid resonant frequencies between 0 and π.

Although stability and frequencies bounded between 0 and π can be enforced using the methods discussed in sect. §3.4.1.1 on page 49, constraints on the region *within* the unit circle is still necessary to enforce *resonant* poles. Therefore, the following general relation between the resonant frequencies and bandwidths to the poles of the

TVAR parameters in eqns. (3.40), (3.41) and (3.43) should be used instead:

$$f_{k,t} = \frac{f_s}{2\pi} \arccos\left(\frac{1 + r_{k,t}^2}{2r_{k,t}} \cos\phi_{k,t}\right) \tag{8.10}$$

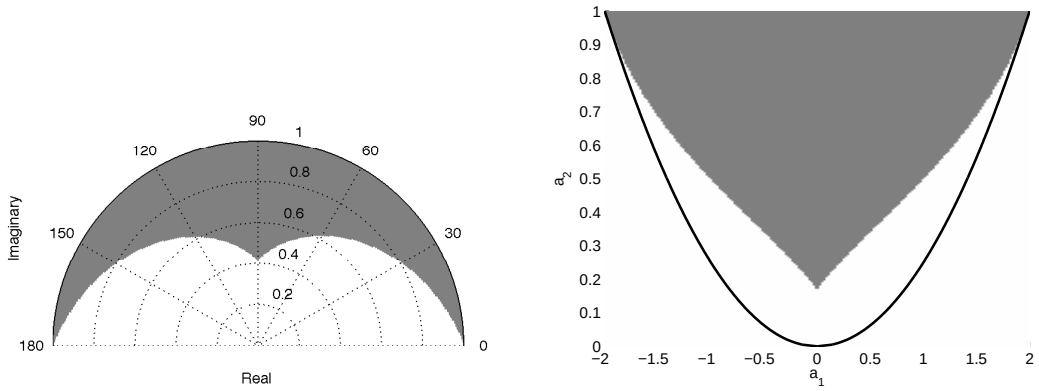
$$B_{k,t} = \omega_u - \omega_\ell \tag{8.11}$$

$$\omega_{\{u,\ell\}} = \arccos\left\{\frac{1}{2r_{k,t}}\left((1 + r_{k,t}^2)\cos\phi_{k,t} \pm (1 - r_{k,t}^2)\sin\phi_{k,t}\right)\right\}. \tag{8.12}$$

However, as only the frequency, $f_{k,t}$ and bandwidth, $B_{k,t}$, are available, four unknowns, i.e., the pole radius and phase, $r_{k,t}$ and $\phi_{k,t}$ respectively, and the upper and lower band-edge frequencies, ω_u and ω_ℓ respectively, need to be solved from two knowns. Clearly, this system is underdetermined and a unique solution does not exist.

Using a general relation valid over the whole unit circle between the resonant parameters and the TVAR coefficients resolves the issue of invalid transformations between the resonant frequency and bandwidth and the TVAR parameters. However, as the frequency response becomes increasingly flat with decreasing pole radius, AR parameters with poles sufficiently close to the origin correspond to magnitude responses where no band-edge frequencies can be identified and a valid 3dB bandwidth does not exist. *Vice versa*, valid resonant frequencies and 3dB bandwidths do not necessarily ensure stable poles as marginally or unstable pole positions close to the unit circle correspond to magnitude responses with well defined peaks. Therefore, a random walk on the resonant frequencies and bandwidths is deemed inappropriate for the modelling the resonant parameters of the PFS model. Therefore, a sampling scheme is necessary that enforces stable parameters corresponding to valid resonant frequencies and bandwidths.

Sect. §8.4 thus investigates the relation between stable AR parameters and valid resonant frequencies and 3dB bandwidths. The region of stable AR parameters corresponding to valid resonant parameters is derived. As this region corresponds to a rather complicated shape in both the parameter space as well as the z-plane (i.e., the pole space), the shape of the valid and stable region of parameters is approximated and approximation accuracy compared for three different parameter spaces. Based on the most accurate approximation, the importance sampling scheme is revised in sect. §8.5 to ensure stable and valid parameters only. Experimental results, comparing the proposed novel source model, amongst others, to the PFS model in sect. §8.4 and the TVAR model in Chap. 7 are discussed in sect. §8.6 and conclusions drawn in sect. §8.7.



(a) *Region of stable poles generating valid 3dB bandwidths.*

(b) *Region of stable parameters generating valid 3dB bandwidths. Parameters underneath black line correspond to the stable parameters illustrated in Fig. 3.15 on page 61.*

Figure 8.4: *Regions corresponding to stable AR parameters with valid resonant frequencies and 3dB bandwidths (grey areas), obtained by evaluating a grid of poles inside the unit circle for eqns.* (3.40) *and* (3.41).

8.4 Admissible regions

Letting the resonant frequency and bandwidth of the resonators evolve according to an uncounstrained random walk can lead to a physically incoherent source model. In order to develop a source model that a) ensures stable parameters and b) valid resonant frequencies and bandwidths , the relationship between $f_{t,k}$, $B_{t,k}$ and $a_{q,t,k}$ in eqns. (3.40) and (3.41) is investigated. A grid of 200×200 poles is generated within the unit circle, i.e., with pole radius $0 \leq r_t \leq 1$ and phase $0 \leq \phi_t \leq \pi$. For each pole, the resonant frequency and 3dB bandwidth in eqns. (3.40) and (3.41) are evaluated. Only if the 3dB bandwidth, B_t, exists and both B_t and the resonant frequency, f_t, lie between 0 and π, the stable pole corresponds to a valid 3dB bandwidth and resonant frequency. As a sanity check, the magnitude response, $|H_t(\jmath\omega)|$, of each pole pair is plotted in order to ensure that spectral peaks of at least 3dB magnitude can be identified. The region of stable poles corresponding to valid resonant frequencies and 3dB bandwidths is found identical for both experiments and displayed as a grey shape in Fig. 8.4a. The corresponding stable AR parameters with valid resonant frequencies and 3dB bandwidths are shown in Fig. 8.4b. Whilst the stable and valid region in the z-plane resembles an hour-glass shape, the stable and valid area in the AR parameter space looks akin to a hybrid shape between a triangle and an ellipse.

Due to the unusual shape of the valid and stable regions in both the pole and parameter space, an exact description of the regional shape is not obvious and cannot

be found straightforwardly. Instead, approximations are required. Sect. §8.4.1 thus examines the approximation of the stable and valid region of AR parameters, whilst sect. §8.4.2 considers the approximation in the pole space. As discussed in sect. §3.4.3 on page 57 ff., the TVAR source model can also be described in terms of the reflection (or PARCOR) coefficients of a lattice structure. PARCOR models are particularly popular models in speech processing applications due to their direct relation to the reflection coefficients of the vocal tract as well as parameters that are guarantee stability of the process. Therefore, sect. §8.4.3 examines the valid and stable region and its approximation in the PARCOR coefficient space.

8.4.1 Admissible regions in parameter space

As illustrated in Fig. 8.4b, the region of AR parameters corresponding to stable coefficients and valid resonant frequencies and 3dB bandwidths of the PFS model resembles a hybrid shape between a triangle and an ellipsoid. One would therefore expect to approximate the boundaries of this shape either by an ellipse or a triangle.

An ellipse centred about the origin with $a_{1,t}$ on the horizontal axis and $a_{2,t}$ on the vertical axis can be described in its well-known canonical form as

$$\left(\frac{a_{1,t}}{\max\{a_{1,t}\}}\right)^2 + \left(\frac{a_{2,t}}{\max\{a_{2,t}\}}\right)^2 = 1. \tag{8.13}$$

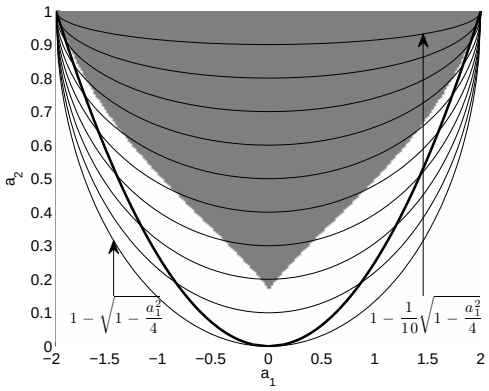
Solving for $a_{2,t}$, eqn. (8.13) can be expressed as

$$a_{2,t} = \max\{a_{2,t}\}\sqrt{1 - \left(\frac{a_{1,t}}{\max\{a_{1,t}\}}\right)^2}. \tag{8.14}$$

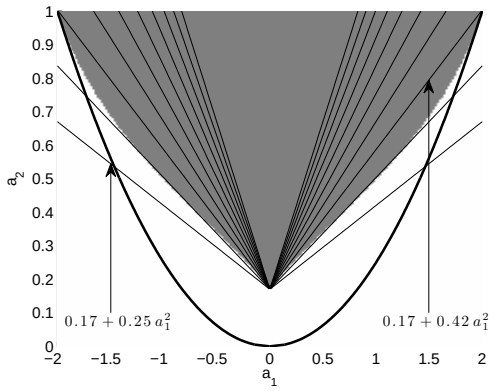
Considering that, from Fig. 8.4b, $a_{1,t}$ is bounded by $-2 \leq a_{1,t} \leq 2$, such that $\max\{a_{1,t}\}^2 = 4$, mirroring the ellipse about the horizontal axis and shifting the origin to 1, an expression for an elliptical approximation of the shape in Fig. 8.4b can be expressed as

$$a_{2,t} = 1 - \max\{a_{2,t}\}\sqrt{1 - \frac{a_{1,t}^2}{4}} \tag{8.15}$$

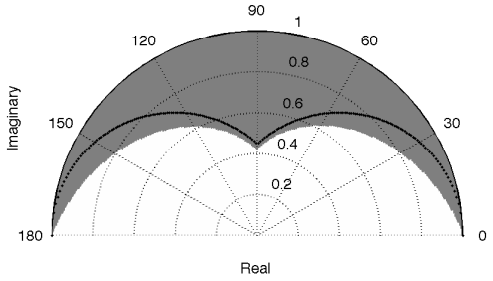
As $\max\{a_{2,t}\}$ is unknown, eqn. (8.15) is plotted in comparison to the stable and valid AR region for $\max\{a_{2,t}\} = 1/10, \ldots, 1$ in Fig. 8.5a. However, the resulting ellipses represent a poor approximation of the area in that either a significant amount of the admissible region is excluded, or parts of the region corresponding to unstable parameters are included in the bounding function. Therefore, an elliptical approximation of the stable and valid region in the AR parameter space seems insufficient.



(a) *Approximation of admissible region in parameter space by ellipses, where* $a_{2,t} = 1 - \alpha\sqrt{1 - a_{1,t}^2/4}$ *for varying parameters* $\alpha = 1, \ldots, \frac{1}{10}$.



(b) *Approximation of admissible region in parameter space by triangles, where* $a_{2,t} = 0.17 + \alpha\sqrt{a_{1,t}^2}$ *for varying* $\alpha = 1/4, \ldots, 1$.



(c) *Approximation in parameter by triangle using* $\alpha = 0.4$ *in pole space.*

Figure 8.5: *Grey areas correspond to regions of stable parameters generating valid 3dB bandwidths. Area underneath thick-lined ellipse in parameter space plots corresponds to the region of stable parameters according to Fig. 3.15.*

Instead of an ellipse, the shape in Fig. 8.4b can be approximated using an isosceles triangle, i.e., in canonical form:

$$a_{2,t} = \begin{cases} \frac{\max\{a_{2,t}\}}{\max\{a_{1,t}\}} a_{1,t} & \text{for } a_{1,t} < 0 \\ -\frac{\max\{a_{2,t}\}}{\max\{a_{1,t}\}} a_{1,t} & \text{for } a_{1,t} < 0 \end{cases} \tag{8.16}$$

The tip of the admissible region is located at approximately $a_{2,t} = 0.17$, such that $\max\{a_{2,t}\} = 1 - 0.17 = 0.83$. The gradient of the triangle stretching from the tip of the valid region to the maximum and minimum of $a_{1,t}$ is therefore $\max\{a_{2,t}\}/\max\{a_{1,t}\} = 0.83/2 \approx 0.41$. Shifting the tip from 0 to 0.17, the triangle that best approximates the desired region is therefore given by

$$a_{2,t} = 0.17 \pm 0.41 a_{1,t}. \tag{8.17}$$

Fig. 8.5b verifies these results by comparing the fit of the triangle in eqn. (8.17) to

triangles with increasing gradients between $\alpha = 1/4, \ldots, 1$. As illustrated, gradients steeper than 0.41 cause the omission of large portions of the valid and stable region of AR parameters, whereas more gentle gradients include unstable AR parameters and coefficients with invalid 3dB bandwidths and resonant frequencies to be included in the area of support. In contrast, the triangle specified in eqn. (8.17) omits the smallest portion of the valid regions and avoids the inclusion of invalid areas.

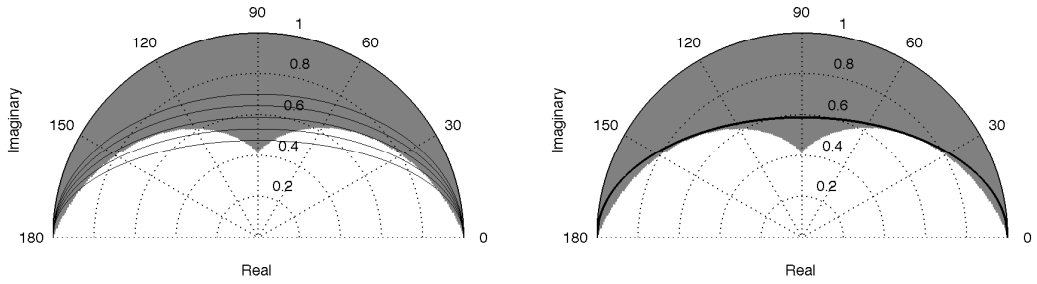
In order to investigate how the the triangular region of support compares to omitted resonances in the frequency spectrum, i.e., poles close to the unit circle, 200 sets of parameters are generated on the boundary lines specified by eqn. (8.17). The corresponding parameters are converted to pole space using eqn. (3.45) and are plotted as a black line in comparison to the region of stable and valid poles in Fig. 8.6b. The region corresponding to eqn. (8.17) in the z-plane is depicted in Fig. 8.5c. From these results it can be seen that the triangular approximation in parameter space captures the region of support in pole space resembling an hour-glass figure well. However, due to the omission of the elliptical sides of the shape in parameter space, the triangular approximation fails to include a considerable portion of poles with radii greater than 0.8 in pole space. The omission of these areas restricts the speech model to resonances corresponding to poles with phases between approximately 30° and 150°. Recalling the plot of the variation of poles with time extracted from a real speech signal in Fig. 3.10b on page 47, poles can be exhibited in speech in the region excluded by the triangular parameter approximation. To avoid to avoid identifiability issues of poles not contained in the valid region of support of the model parameters, it is therefore desirable to specify a more accurate approximation of the valid areas in the pole space.

8.4.2 Admissible regions in the z-plane

In order to reduce the number of resonant and valid poles excluded from the approximated region of support, the valid and stable region can be approximated directly in the z-domain rather than the AR parameter space. Again, due to the shape of the region of support in Fig. 8.4a, an exact description of the boundaries seems non-obvious. Ellipses are thus used instead to approximate the region. Similar to sect. §8.4.1, an ellipse centred about the origin with $\max\{r_t\} = 1$ can be expressed according to eqn. (8.14) as

$$\phi_t = \max\{\phi_t\}\sqrt{1 - r_t^2}, \tag{8.18}$$

where the imaginary part is normalised between $0 \leq \phi_t \leq 1$. As the maximum of the unsupported region in Fig. 8.4a lies at approximately 0.5, the ellipses corresponding to $\max\{\phi_t\} = 0.475, \ldots, 0.7$ are plotted in Fig. 8.6a. As the isolated plot in Fig. 8.6b



(a) *Ellipses approximating the admissible region for* $\alpha = 0.46$ *(bottom) to* $\alpha = 1$ *(top).*

(b) *Admissible region* $\max\{0.5875\}$

Figure 8.6: *Grey areas correspond to regions of stable parameters generating valid 3dB bandwidths.*

illustrates, the best approximation is achieved for $\max\{0.5875\}$, omitting the least portion of the valid region and avoiding the inclusion of invalid poles.

As compared to the triangular parameter approximation in Fig. 8.5c, only a small region of poles with $r_t > 0.8$ is excluded. However, the elliptical pole approximation in Fig. 8.6b fails to model the lobe between $120 \leq \phi_t \leq 60$ and $0.4 \leq r_t \leq 0.6$. Nonetheless, the magnitude responses are comparatively flat in this region. The flatness of the magnitude response of the excluded lobe area is demonstrated in Fig. 8.7 where the responses for poles with radius $r_t = 0.5$ and phases between 80 and 100 degrees are compared to the response with poles of $r_t = 0.9$ for the same phase values. As illustrated, the peak of the magnitude response for $r_t = 0.5$ is approximately 3dB high, whereas the response for $r_t = 0.9$ reveals peaks of up to 20dB height.

Therefore, the poles in the excluded lobe only cause a negligible contribution to the frequency response. As only only a small region of resonant poles with radii above 0.8 is omitted in the elliptical approximation of the region of support in pole space as compared to the triangular approximation in parameter space, the elliptical pole approximation in Fig. 8.6b using eqn. (8.18) proves a more practical option than the triangular parameter approximation in Fig. 8.5c using eqn. (8.17).

As the approximated region of support in Fig. 8.6b still excludes a small portion of resonant poles between $0 \leq \phi_t \leq 40$ and $180 \leq \phi_t \leq 140$. Ideally, an approximation of the region of support should thus be identified that only excludes the central lobe of the hour-glass shape, whilst including any valid areas close to the unit circle.

As discussed in sect. §3.4.2 on page 52, the AR parameters are not only related

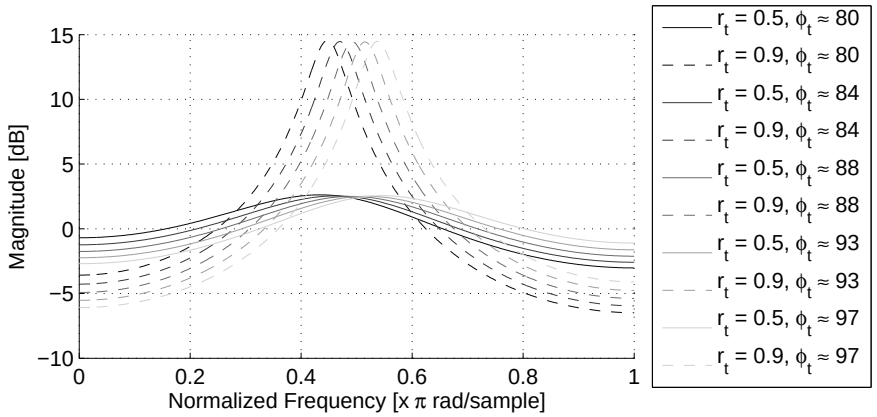


Figure 8.7: *Comparison of magnitude responses for poles with radius* $r_t = 0.5$ *(solid) and* $r_t = 0.9$ *(dashed) for phases between* 80 *and* 100 *degrees.*

to the complex poles in z-plane, but can be interpreted as PARCOR coefficients of a lattice structure. An investigation of the region of support in the PARCOR space thus appears as an interesting alternative to the discussions in pole and parameter space in sects. §8.4.1 and §8.4.2.

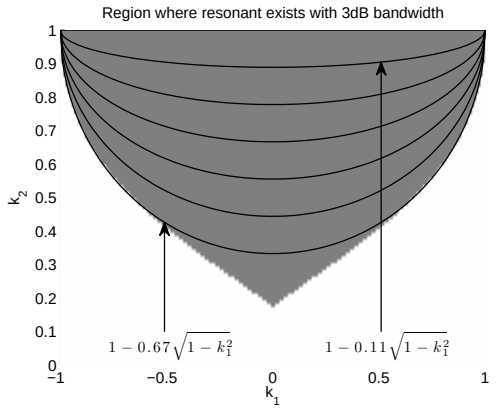
8.4.3 Admissible regions in PARCOR space

Instead of parameterising the TVAR model in terms of AR coefficients using a direct-form infinite impulse response (IIR) structure, the model can be represented by a lattice IIR structure and hence parameterised in terms of the reflection coefficients describing the lattice. As shown in sect. §3.4.2.1 on page 54, the reflection coefficients of the lattice structure correspond to so-called PARCOR coefficients, describing the relation between the forward and backward lattice structure. This description is directly related to the relation between the propagated and reflected sound waves at junctions in the acoustic tube, such that the reflection, or PARCOR, coefficients of the lattice structure are equivalent to the reflection coefficients of the vocal tract transfer function (see sect. §3.4.2.2). Due to this physical interpretation, the reparameterisation of the TVAR speech model in terms of PARCOR coefficients is thus extensively applied in various speech processing applications [107, 142, 144, 145].

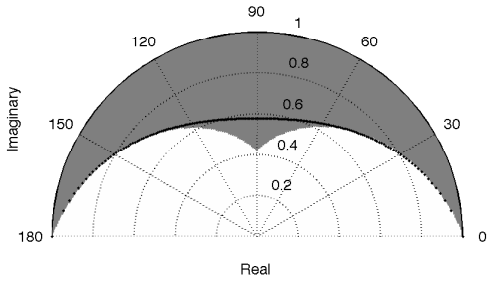
Recalling the discussion in sect. §3.4.2 on page 52, the TVAR parameters are related to the PARCOR coefficients for a second-order model via eqns. (3.28a) and (3.28b) on page 54, i.e.,

$$a_{1,t} = \psi_{1,t}\,(1 + \psi_{2,t}) \tag{8.19}$$

$$a_{2,t} = \psi_{2,t}. \tag{8.20}$$



(a) *Admissible region*



(b) *Region of stable parameters generating valid 3dB bandwidths. Parameters underneath black line correspond to the stable parameters illustrated in Fig. 3.15 on page 61.*

Figure 8.8: *Grey areas correspond to regions of stable parameters generating valid 3dB bandwidths.*

where $\psi_{1,t}$ and $\psi_{2,t}$ are the two reflection (or PARCOR) coefficients of the lattice structure. Similar to sect. §8.4.2, the area of stable AR parameters corresponding to valid resonant frequencies and 3dB bandwidths can therefore be reflected into the PARCOR coefficient domain using eqn. (3.28). The resulting region is shown as a grey shape in Fig. 8.8a. This shape strongly resembles the region of support in AR parameter domain in Fig. 8.4b and, again, resembles a hybrid between a triangle and ellipse. Contrary to the shape in AR parameter domain, however, the body of the shape appears more rounded and voluminous. An ellipse using $\max\{\psi_{1,t}\} = 1$ is therefore fitted to the region of support in PARCOR parameter space, where

$$\psi_{2,t} = 1 - \max\{\psi_{2,t}\}\sqrt{1-\psi_{1,t}^2}. \tag{8.21}$$

The ellipses for $\max\{\psi_{2,t}\} = 0, \ldots, 2/3$ are shown in Fig. 8.8a as black lines. It can be seen that $\max\{\psi_{2,t}\} = 2/3$ delivers the most accurate approximation of the region of support, omitting the triangular central lobe of the shape without including any invalid PARCOR coefficients.

The elliptical PARCOR parameter approximation is transformed into the z-plane to draw conclusions about the omission of resonant poles as discussed in sects. §8.4.1 and §8.4.2. The elliptical PARCOR approximation is thus compared to the valid and stable region of support in pole space in Fig. 8.8b. Compared to Figures 8.5c and 8.6b, the PARCOR approximation omits the central lobe between $60 \leq \phi_t \leq 120$ and $0.4 \leq r_t \leq 0.6$ like the elliptical pole approximation in sect. §8.4.2. However, unlike either the pole or the AR parameter approximation, the PARCOR approximation does

not exclude any resonant poles close to the unit circle. Therefore, the approximation in PARCOR parameter space provides the most accurate approximation of the region of support in the z-plane as compared to the approximation in pole space or AR parameter space.

Hence, the sampling scheme in sect. §8.3 should be reparameterised in terms of valid PARCOR coefficients. As an expression for boundaries on the PARCOR parameters is available through eqn. (8.21) as illustrated in Fig. 8.4.3, PARCOR coefficients corresponding to valid frequencies and stable parameters can be enforced by *transforming* the samples into valid and stable variates. Hence, instead of sampling the resonant frequencies and bandwidths from an unconstrained random walk as proposed in [111] and discussed in sect. §8.3, the PARCOR coefficients are sampled from a random walk and projected into the boundaries enforcing stable parameters and valid frequencies. Parameterisation of the model in terms of PARCOR coefficients and enforcing samples from the valid region of support in Fig. 8.8a hence avoids instability of the TVAR parameters, as well as invalid resonant frequencies and 3dB bandwidths.

8.5 Reparameterisation of the source for stability

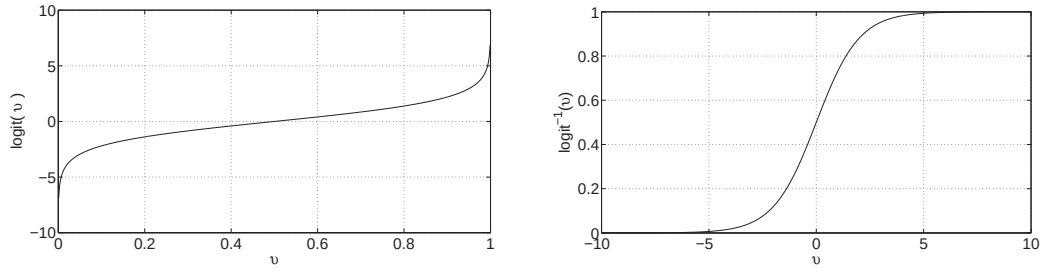
Rather than modelling the resonant frequency and bandwidth of the resonators as a random walk, the PARCOR coefficients of the resonator are modelled as

$$\boldsymbol{\psi}_t = \boldsymbol{\psi}_{t-1} + \boldsymbol{\Sigma}_{\boldsymbol{\psi}_t} \mathbf{r}_{\boldsymbol{\psi}_t} \qquad \mathbf{r}_{\boldsymbol{\psi}_t} \sim \mathcal{N}(2,\, 1) \tag{8.22}$$

where $\boldsymbol{\psi}_t = \begin{bmatrix} \psi_{1,t,k} & \psi_{2,t,k} \end{bmatrix}^T$ and $\boldsymbol{\Sigma}_{\boldsymbol{\psi}_t} = \operatorname{diag}\begin{bmatrix} \sigma_{\psi_{1,t,k}} & \sigma_{\psi_{2,t,k}} \end{bmatrix}$ is the covariance of the coefficients. In order to ensure that the samples lie within the region of support illustrated in Fig. 8.8a samples drawn from eqn. (8.22) that lie outside of the region need to be reflected back into the valid area. Similar to enforcing stable AR parameters as discussed in sect. §3.4.1.1, valid and stable PARCOR coefficients can be obtained by the introduction of an indicator function that accepts the generated sample if it is valid and stable or rejects it otherwise. However, as in the case for AR parameters, the rejection of samples can lead to the depletion of computational effort and numerical issues.

Instead of rejecting unstable or invalid samples, it is proposed to utilise bounded functions to transform *all* samples into a bounded area. Any bounded function, e.g., inverse trigonometric functions such as the arctan or arcsin, can be used. In this chapter, the inverse logit is utilised. The logit of a number $0 \leq \upsilon \leq 1$ is defined as

$$\chi = \operatorname{logit}(\upsilon) \triangleq \ln\left\{\frac{\upsilon}{1-\upsilon}\right\} \tag{8.23}$$



(a) *logit of* $0 \leq \upsilon \leq 1$ **(b)** *logit*$^{-1}$ *of* $-10 \leq \upsilon \leq 10$

Figure 8.9: *Definition of the logit for creating samples bounded between two values*

such that $-\infty \leq \chi \leq \infty$ (see Fig. 8.9a). The inverse logit-function is thus bounded between 0 and 1 for any variable $-\infty \leq \chi \leq \infty$, where

$$\upsilon = \text{logit}^{-1}(\chi) = \frac{1}{1 + e^{-\chi}} \tag{8.24}$$

as shown in Fig. 8.9b. Therefore, any unbounded variable, $\chi \in \mathbb{R}$, can be bounded between 0 and 1 by applying eqn. (8.24).

The stable and valid area of PARCOR coefficients in Fig. 8.8a are constrained between $-1 \leq \psi_{1,t} \leq 1$ and $1 \leq \psi_{2,t} \leq 1 - 2/3\sqrt{1 - \psi_{1,t}^2}$, whereas the inverse logit is defined between 0 and 1. Eqn. (8.24) should thus be modified to account for any lower limit in order to accommodate the boundaries set by the region of support in PARCOR parameter space. Shifting and scaling eqn. (8.24) by some lower limit, α, the inverse logit can be bounded between $\alpha \leq \upsilon \leq 1$ via

$$\upsilon = \alpha + \frac{1 - \alpha}{1 + e^{-\chi}} = \alpha + (1 - \alpha)\,\text{logit}^{-1}(\chi).$$

where $\alpha \in \mathbb{R}$.

The boundary conditions for stable and valid PARCOR coefficients can thus be enforced by firstly drawing an auxiliary sample of the two PARCOR coefficients according to eqn. (8.22), i.e.,

$$\widehat{\psi}_{1,t,k} = \psi_{1,t-1,k} + \sigma_{\psi_{1,t,k}}\, r_{\psi_{1,t,k}} \tag{8.25}$$

$$\widehat{\psi}_{2,t,k} = \psi_{2,t-1,k} + \sigma_{\psi_{2,t,k}}\, r_{\psi_{2,t,k}} \tag{8.26}$$

and secondly enforcing the PARCOR coefficients to lie between $-1 \leq \hat{\psi}_{1,t,k} \leq 1$ and $\varphi(\psi_{1,t,k}) \leq \psi_{2,t,k} \leq 1$ where $\varphi(\psi_{1,t,k}) = 1 - \frac{2}{3}\sqrt{1 - \psi_{1,t,k}^2}$ is the elliptical approximation of the region of support. Hence, the transformed PARCOR coefficients are

for $t > \max\{P, Q\}$ **do**
  for $i = 1, \ldots, N$ **do**
    for $k = 1, \ldots K$ **do**
1 Importance sampling of the resonant gain, $g_{t,k}^{(i)}$, (eqn. (8.9c)) and auxiliary PARCOR parameters, $\hat{\boldsymbol{\psi}}_{t,k}^{(i)}$, (eqn. (8.25));
2 Transformation of auxiliary coefficients, $\hat{\boldsymbol{\psi}}_{t,k}^{(i)}$, to the valid and stable region in PARCOR coefficients space (eqn. (8.27));
3 Transformation of PARCOR coefficients, $\boldsymbol{\psi}_t^{(i)}$, to source parameters, $\mathbf{a}_{t,k}^{(i)}$ (eqn. (3.28));
    end
4 Importance sampling of measurement noise factor $\phi_{w_t}^{(i)}$ (eqn. (6.5));
5 Kalman filter prediction of $\boldsymbol{\mu}_{t|t-1}^{(i)}$, $\boldsymbol{\Sigma}_{t|t-1}^{(i)}$ (eqns. (6.21a) and (6.21b)).;
6 Kalman filter estimation of $\boldsymbol{\mu}_{\mathbf{b},t}^{(i)}$ and $\boldsymbol{\Sigma}_{\mathbf{b},t}^{(i)}$ (eqn. (6.22));
7 Kalman filter correction of $\boldsymbol{\mu}_{t|t}^{(i)}$, $\boldsymbol{\Sigma}_{t|t}^{(i)}$ (eqns. (6.21c) and (6.21d));
8 Evaluation of weights $w_t^{(i)}$ (eqns. (6.38), (6.32));
  end
9 Normalization of importance weights;
10 Resampling;
11 Computation of particle average:

$$\widehat{\mathbf{x}}_t = \sum_{i \in \mathcal{N}} \hat{\boldsymbol{\mu}}_{t|t}^{(i)} \qquad \widehat{\boldsymbol{\theta}}_t = \sum_{i \in \mathcal{N}} \boldsymbol{\theta}_{0:t}^{(i)}, \qquad \widehat{\mathbf{b}} = \sum_{i \in \mathcal{N}} \boldsymbol{\mu}_{\mathbf{b},t}^{(i)}.$$

end

Algorithm 8.1: RBPF for parallel formant synthesizer model with PARCOR coefficient sampling

expressed as

$$\psi_{1,t,k} = -1 + \frac{2}{1 + \exp\left\{-\widehat{\psi}_{1,t,k}\right\}} \tag{8.27a}$$

$$\psi_{2,t,k} = \alpha + \frac{1-\alpha}{1 + \exp\left\{-\widehat{\psi}_{2,t,k}\right\}}. \tag{8.27b}$$

Therefore, two issues with the sampling scheme in sect. §8.3 are resolved: Firstly, as $\boldsymbol{\psi}_{0:t}$ is enforced to lie in the stable and valid region of support shown in Fig. 8.4.3, resonant frequencies are ensured to lie between 0 and π, whilst the corresponding peaks are ensured to have valid 3dB bandwidths. Secondly, only one transformation from the PARCOR to the AR parameter space is necessary, rather than several transformations between the resonant frequencies and bandwidths, the AR parameter and pole space. The proposed sampling scheme therefore facilitates a sampling scheme avoiding multiple transformations between parameter spaces and ensuring stable pa-

rameters that correspond to valid frequencies and bandwidths.

The proposed parameter model is therefore implemented in the PFS model to parameterise the resonator circuits. The resulting novel PFS model is implemented in the RBPF as summarised in Alg. 8.1. Experimental results are presented in the following section for speech data and compared to the estimated obtained using the dynamic TVAR parameter model in Chap. 7.

8.6 Experimental results

8.6.1 Speech data

In order to evaluate the performance of the RBPF using the proposed PFS model in Alg. 8.1 and compare it to the dynamic TVAR parameter model in Chap. 7, the experiment in sect. §7.4.2 on page 138 is repeated. Using the the PFS model, the RBPF is thus executed using 1000 particles and $K = 6$ resonators for the speech signal at $f_s = 4$kHz distorted by WGN and the 8-th order gramophone horn model.

The resulting distortion measures are summarised and compared to the results of the dynamic TVAR parameter model in Table 8.1. The RBPF using the PFS model achieves a segmental signal-to-reverberant component ratio (SRR) of 5.80dB and an log-spectral distortion (LSD) of 1.22dB. Compared to the observed signal SRR of -6.59dB and LSD of 1.67dB, a SRR improvement of 12.39dB and a LSD improvement of 0.44dB is achieved. Compared to the estimates obtained using the dynamic TVAR parameter model, the PFS performs 2.83dB better in terms of the segmental SRR and 0.13dB better in terms of the LSD.

Similar to the dynamic TVAR model, the RBPF using the PFS model recovers high-energy components in the base-band between $0 - 200$Hz as illustrated by the spec-

	Segmental SRR	LSD
Observed signal	-6.59	1.67
TVAR parameter model	2.97	1.35
PFS model in [111]	2.12	1.31
PFS model	5.80	1.22
Improvement over observed signal	12.39	0.44
Improvement over TVAR parameter model	2.83	0.13
Improvement over PFS model in [111]	3.68	0.09

Table 8.1: *Distortion measures for speech data comparing the RBPF estimate and observed signal distortion.*

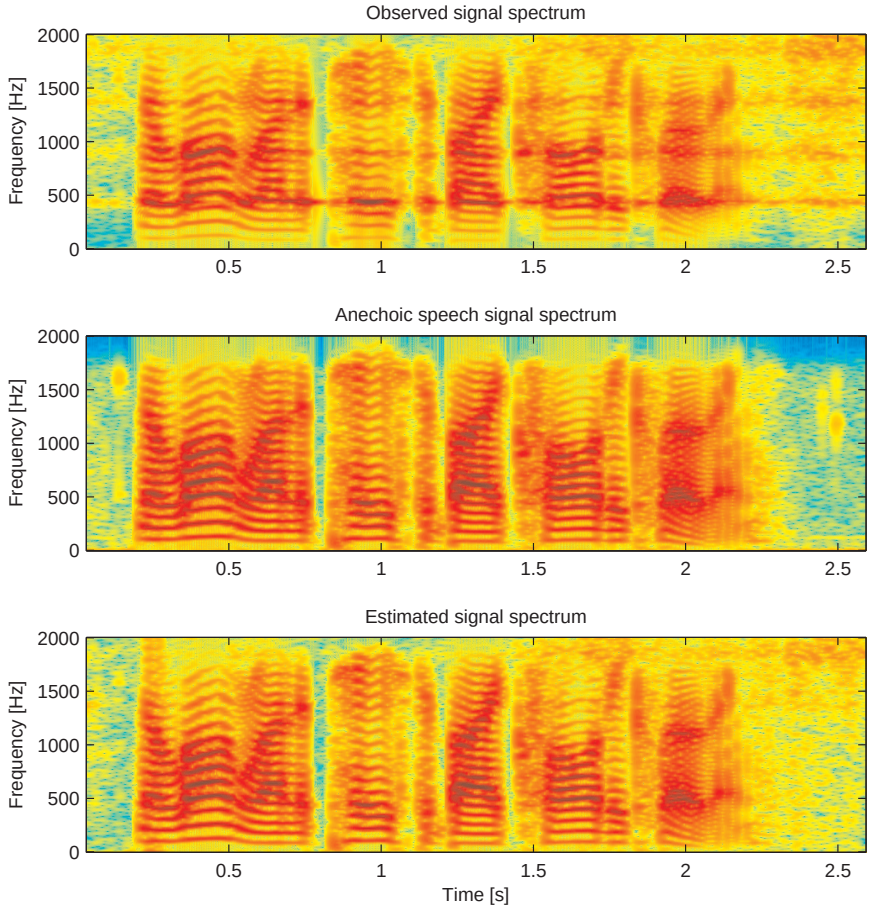


Figure 8.10: *Spectrograms of the anechoic, reverberant, and estimated speech signal, "In the long run, it pays to buy quality clothing." at* $f_s = 4kHz$*. Color scale from dark red (high energy) to dark blue (low energy) indicates the dynamic range of the signal between* $1dB$ *and* $-7dB$*.*

trogram in Fig. 8.10. Whilst energy introduced through reverberation in the silent period at the end of the sentence are slightly reduced as compared to the observed signal, high-energy frequency tracks are reconstructed, e.g., around 2s and between $1 - 1.3kHz$. An audio sample demonstrating that the audio quality is improved over both the reverberant observed signal as well as the dynamic TVAR source model estimates in Chap. 7 can be found on the attached compact disc (CD) in the folder '*Chapter 5-6 - TVAR and PFS model*'.

In order to demonstrate improvement in dereverberation performance over the PFS model by Beierholm and Winther, the experiment is repeated for the PFS model in [111] as discussed in sect. §8.3. Here, importance samples of the resonant frequency and bandwidth are drawn directly from the random walk in eqn. (8.9) and related to the TVAR parameters of the speech model using eqn. (3.36). This approach leads

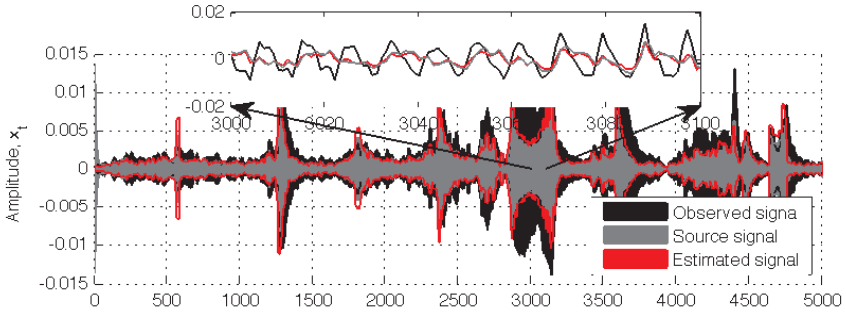
to unbounded samples of the resonant frequencies and bandwidths and hence the TVAR parameters. However, it was found that a direct application of this approach leads to computational errors due to importance samples corresponding to unstable TVAR parameters. In order to ensure TVAR parameters within the unit circle, unstable choices are reflected back into the unit circle as discussed in sect. §3.4.1.1 on page 49. Nonetheless, it should be noted that resonant frequencies between $0 \leq f_{t,k} \leq \pi$ and $0 \leq B_{t,k} \leq \pi$ can only be enforced by choosing low variances on the corresponding random walk. However, formant frequencies of the vocal tract can vary rapidly in frequency bands. A low variance on the resonant frequency of the PFS model does not reflect this rapid variation of formant frequencies.

Application of the PFS model by Beierholm and Winther [111] in the RBPF framework to the speech signal described above for 1000 particles and 6 resonators, initialised at $f_{t,k} = B_{t,k} = \pi/2$ results in a segmental SRR of the estimated signal of 2.12dB and a LSD of 1.31dB. An audio sample demonstrating that the audio quality of the PFS model according to [111] can be found on the attached CD in the folder '*Chapter 5-6 - TVAR and PFS model*'. As summarised in Table 8.1, the PFS-PARCOR model proposed in this chapter thus leads to an improvement of 2.83dB in the segmental SRR and 0.09dB in the LSD compared to the PFS model proposed in [111]. Furthermore, both valid resonant frequencies and bandwidths as well as stable TVAR parameters are enforced using the PFS model proposed in this chapter.

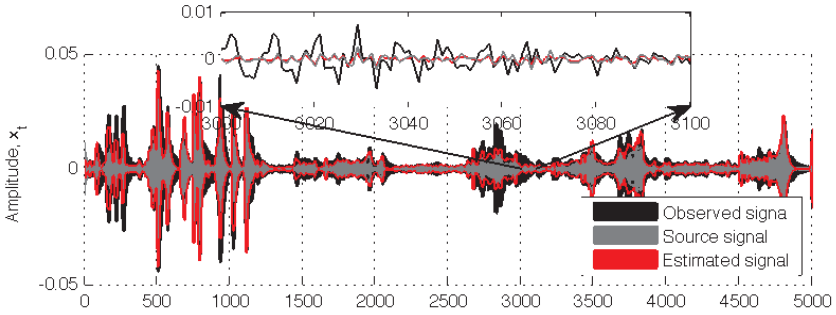
8.6.2 Investigation of performance for different phoneme types

Next, the SRR improvement of the PFS model is compared to the dynamic TVAR parameter model for different phoneme types. Hence, the experiment in sect. §7.4.3 on page 141 over the databases of phoneme types is repeated for the PFS model. The results in terms of the distortion measures, i.e., the SRR and LSD, are summarised in Table 8.2 and compared to the results for the dynamic TVAR parameter model in Table 7.3 on page 142.

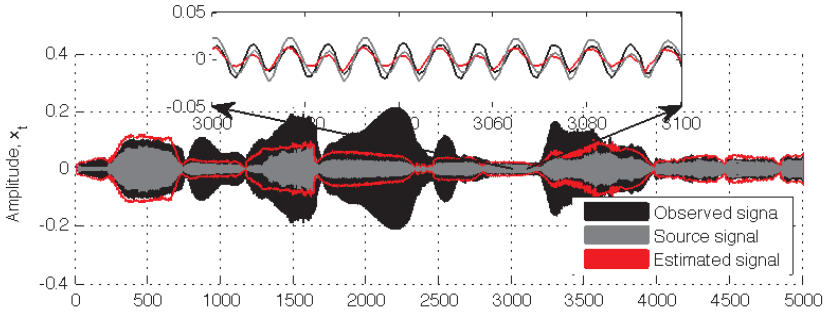
From these results, it is evident that an overall improvement is achieved by using the PFS model over the dynamic TVAR parameter model. For instance, the SRR of fricatives was improved by 3.9dB over the SRR of the dynamic TVAR parameter model, whilst stop consonants are improved by 3.22dB over the dynamic TVAR model. Signal plots in Figures 8.12a and 8.12b compare the results of the dynamic TVAR and PFS model for fricatives and stop consonants. These figures highlight more rapid adjustment to changes and improved accuracy in modelling the signal amplitude of the underlying data by the PFS model. For fricatives, Fig. 8.12a highlights the accurate modelling of both the PFS and TVAR model for the change from a sig-



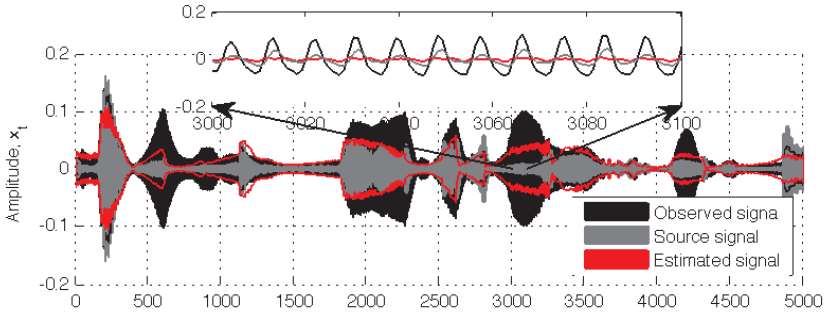
(a) *Fricative sounds*



(b) *Stop consonants*

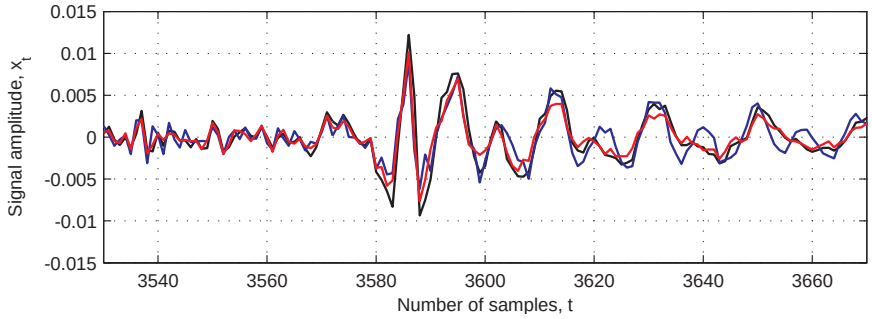


(c) *Vowels*

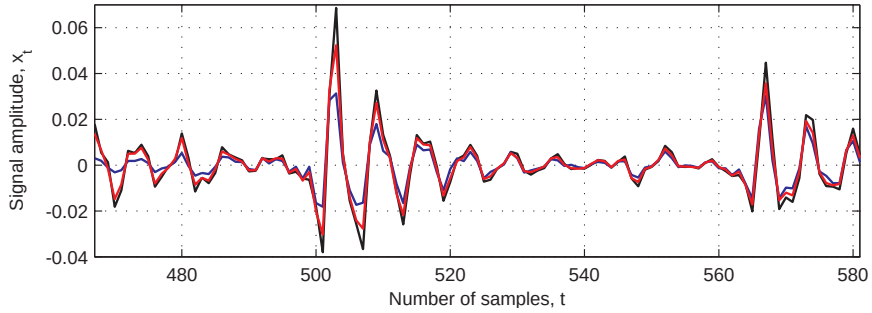


(d) *Semivowels*

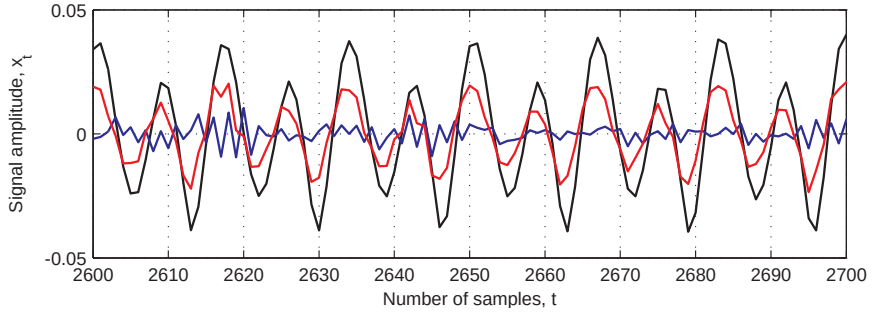
Figure 8.11: *Comparison of the anechoic source, observed, and estimated signals for four speech sequences containing either fricatives, stop consonants, vowels, or semivowels.*



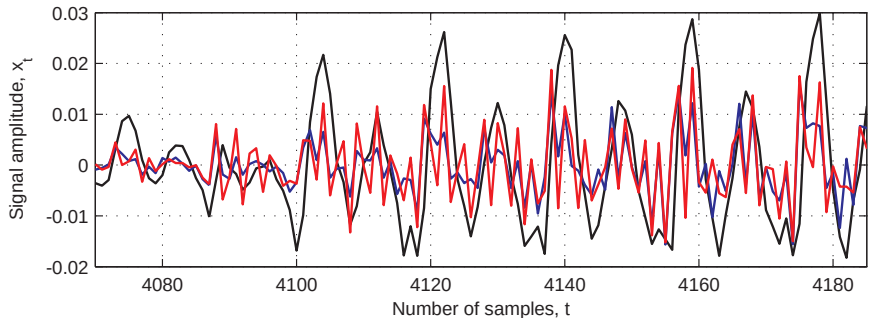
(a) *Transient segment in a fricative*



(b) *Abrupt changes in consonants*



(c) *Harmonic segment in a vowel*



(d) *Abrupt changes in semivowels*

Figure 8.12: *Comparison of the dynamic TVAR parameter model (blue) and the PFS model (red) for different speech phonemes (black)*

Phoneme type		Segmental SRR	LSD
Fricatives	Observed signal	−3.11	1.19
	TVAR parameter model	2.14	0.82
	PFS model	6.04	0.71
	Improvement over TVAR model	3.9	0.11
	Improvement over observed signal	9.15	0.48
Stop consonants	Observed signal	−2.65	1.261
	TVAR parameter model	4.16	1.268
	PFS model	7.38	1.012
	Improvement over TVAR model	3.22	0.256
	Improvement over observed signal	1.03	0.249
Vowels	Observed signal	−2.70	1.17
	TVAR parameter model	−0.88	1.79
	PFS model	3.36	1.35
	Improvement over TVAR model	4.24	0.44
	Improvement over observed signal	6.06	−0.18
Semivowels	Observed signal	−4.19	1.28
	TVAR parameter model	0.42	1.74
	PFS model	−1.49	1.83
	Improvement over TVAR model	−1.91	−0.09
	Improvement over observed signal	2.7	−0.55

Table 8.2: *Distortion measures for speech data comparing the RBPF estimate and observed signal distortion.*

nal segment resembling WGN up to $t = 3580$ and the change to an almost periodic segment thereafter. For stop consonants (see Fig. 8.12b), the PFS model captures the abrupt changes in amplitude well. This can be particularly well seen at approximately $t = 500$. In contrast, the TVAR captures less abrupt changes (as seen at approximately $t = 565$), but does not capture the variation in amplitude at $t = 500$ as well as the PFS model. Similarly, the PFS model approximates the sinusoidal properties of vowels in Fig. 8.12c significantly better than the TVAR model. Due to the very short duration and rapid variation of characteristics of semivowels, both the PFS and TVAR model struggle to reconstruct the anechoic signal segment in Fig. 8.12d.

Vowels, whose harmonicity was insufficiently captured by the dynamic TVAR model, are improved by 4.24dB to achieve an SRR of 3.36dB using the PFS model. This improvement is well illustrated by the time-domain plot of the vowel approximation in Fig. 8.11c and in Fig. 8.11c: the PFS model captures the sinusoidal approximates the sinusoidal variation of the underlying speech segment significantly better than the TVAR model. In contrast, the TVAR model resembles the mean of the signal in Fig. 7.10c rather than harmonic behaviour of the anechoic speech signal. A direct

comparison between the PFS and TVAR model for harmonic components extracted from the vowel segment is illustrated in Fig. 8.12c. In this figure, in can be seen that the TVAR model attempts to approximate the sinusoidal speech envelope by a variation resembling turbulent noise. However, the PFS model captures approximates the harmonic signal albeit slightly attenuated.

Semivowels, however, are modelled marginally better by the dynamic TVAR parameter model. Nonetheless, even the dynamic TVAR model only achieves a SRR of 0.42dB. Issues in modelling semivowels can be due to the rapid and sudden variation in what appears as "pulses" of sinusoidal functions in the signal amplitude as illustrated in Fig. 8.11d. A comparison of both models with a speech segment of the semivowel is shown in Fig. 8.11d, highlighting inaccurate approximation of the underlying speech phoneme by either model. Although the vowels in Fig. 8.11c due to their nature as voiced phonemes exhibit strong periodic and harmonic components as well, sinusoidal segments have a longer duration and smoothen out towards the end of the phoneme. Instead, semivowels seem to exhibit sharp short blocks of combinations of sinusoids and turbulent noise, starting and ending abruptly.

8.7 Discussion

In this chapter, a source signal model based on parallel formant synthesis was introduced, in order to improve upon performance of the dynamic TVAR model in Chap. 7 for modelling fricatives and vowels. Three to five resonator circuits are connected in parallel to model the three to five formants in human speech. PFSs in the speech *synthesis* literature are parameterised by the resonant frequency and bandwidth of each resonator circuit. In order to synthesise a speech sound, the frequency and bandwidth are assumed known, controlling the amplitude control of the resonator. As resonator circuits can be expressed as second-order TVAR models, the frequency and bandwidth of the circuit can be related to the TVAR parameters of the source model.

In speech estimation, as opposed to speech synthesis, source signals are *approximated* without prior knowledge of the resonator frequency and bandwidth. Therefore, a model on the resonator parameters is necessary in order to model the dynamic of the source signal. Following the reasoning in Chap. 7, it seems tempting to model the resonator frequency and bandwidth, rather than the source parameters, as a random walk. However, the frequency is limited between 0 and π and the resulting spectral peak must be at least 3dB high in order to extract a valid bandwidth. An unconstrained random walk, however, does not enforce these boundaries.

This chapter therefore investigated the AR regions corresponding to stable poles generating valid frequencies and bandwidths of the resonator. It was shown that these regions can be best approximated in the PARCOR space, i.e., by representing the AR source model as a lattice structure. PARCOR coefficients are associated with beneficial properties, such as a direct connection to the reflection coefficients in the acoustic tube model, or natural boundedness in a region generating stable AR parameters. PARCOR representations of AR models are therefore particularly popular in the speech processing community. Using the PARCOR representation, it was proposed to let the PARCOR coefficients evolve according to a random walk and reflecting the resulting parameters into a space corresponding to stable AR parameters and valid frequencies and bandwidths.

Experiments were presented evaluating this model for speech signals and comparing to the results of the dynamic TVAR parameter speech model. It was demonstrated that the performance of the RBPF for fricatives, stop consonants and vowels can be significantly improved by using the PFS model.

Chapter 9

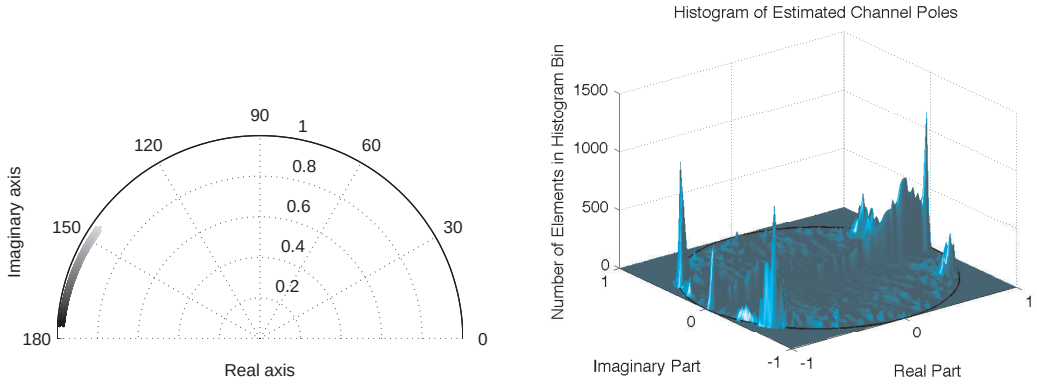
Blind dereverberation of speech from a moving speaker

9.1 Introduction

The assumption of a static, time-invariant room impulse response (RIR) made in Chaps. 6 and 8 is appropriate in scenarios where the source-sensor geometry is not rapidly varying, e.g., for hands-free kits in a car cabin or in work environments where the user is seated in front of a computer terminal. However, users of hands-free conference telephony equipment or wearers of hearing aids usually wish to be able to move freely around a room. By moving with $1\,\mathrm{m/s}$ throughout a room, a distance of 50mm is covered within 50ms, sufficient to render any assumption of a time-invariant acoustic impulse response (AIR) invalid.

Dereverberation of speech for moving speakers is therefore an important but also inherently difficult problem and has received little attention in the research thus far. Approaches that address this problem can be found in [57,64,213]. This chapter therefore proposes an extension of the *static* channel model in Chaps. 6 and 8 to a *time-varying* channel, modelling the variation of the RIR with changing source-sensor positions.

Assuming an AR source model and a time-varying all-pole (TVAP) channel model, both the poles of the source and channel vary with time. When estimating the source and channel poles from the frequency response of the reverberant signal it therefore seems difficult to differ between time-varying source and channel poles. This chapter demonstrates using simulated and measured RIRs that the channel poles vary slowly and smoothly with time. In contrast, the poles of speech vary relatively rapidly with



(a) *Slowly varying channel parameters.* **(b)** *Histogram over the extracted poles.*

Figure 9.1: *Figure illustrating the extraction of slowly varying poles.*

time (see sect. §3.4 on page 48). The underlying idea of this chapter is that the channel can be identified from the rapidly varying source due to the slowly varying channel poles.

This idea can be illuminated using the following example: consider the speech signal "In the long run it pays to buy quality clothing" at $f_s = 4$kHz distorted by the second-order slowly varying channel with poles varying along the unit circle as illustrated in Fig. 9.1a. The distorted observed signal is modelled as an all-pole filter. I.e., assuming $Q = 15$ speech poles and $P = 2$ channel poles, $Q + P = 17$ poles are extracted over a sliding window of block size 100 from the observed signal by solving the Yule-Walker equations using the Levinson-Durbin recursion. The histogram is then taken over all windows as shown in Fig. 9.1b. The histogram exhibits high peaks at the locations of the moving channel poles, whereas the speech poles mostly smear over the bottom of the unit circle. Therefore, due to the slow variation of the channel poles, their pole positions can be identified from the response of the observed signal.

Based on this idea, this chapter investigates the variation of the channel parameters characterising simulated image-source method (ISM) responses and measured RIRs. Based on the results a time-varying channel model is proposed that models the time-varying channel parameters as a linear combination of known time-varying basis functions with unknown time-invariant channel as discussed in, e.g., [128,214–218]. A similar channel model was previously proposed and shown effective for blind dereverberation of moving speakers using Markov chain Monte Carlo (MCMC) methods in [1] The model can be easily integrated in the observation space of the RBPF in eqn. (6.9) on page 115. Experimental results on synthetic and speech data are presented.

This chapter is therefore structured as follows: Sect. §9.2 introduces the extension of the static channel in eqn. (6.9) to time-varying channel parameters. The variation of realistic RIRs and their characterising channel parameters with changing source-sensor positions is investigated in sect. §9.3. Sect. §9.4 examines the approximation of the time-varying channel parameters by polynomial functions. Based on the results, sect. §9.5 introduces the TVAP model using the linear combination of basis functions and discusses its implementation in the RBPF. Experimental results on synthetic and speech data are presented in sect. §9.6 and conclusions drawn in sect. §9.7.

9.2 Non-stationary channel model

The RIR describing the acoustic channel between the source and sensor changes with varying source-sensor geometries. Therefore, as the source changes position with time, the RIR changes with time. Hence, the channel parameters characterising the RIR are time-varying. Therefore, the linear time-invariant (LTI) channel model in eqn. (4.15) on page 77 can be extended to accommodate moving speakers by extension to a linear time-varying (LTV) channel model, i.e.,

$$y_{m,t} = \sum_{p \in \mathcal{P}} b_{m,p,t}\, y_{m,t-p} + x_t + \sigma_{m,w_t}\, w_{m,t}, \tag{9.1}$$

where $\{b_{m,p,t} : m \in \mathcal{M},\, p \in \mathcal{M}\}$ are the time-varying channel parameters. Therefore, the problem of modelling the RIR between varying source and sensor positions reduces to modelling the time-varying all-pole parameters, $\{b_{m,p,t}\}$.

Due to the limited attention moving RIRs have received until recently, specification of time-varying channel models is, to a certain extent, an open question and is partially constrained by the availability of tractable parameter estimation techniques. Therefore, the RIR variation for moving speakers should be investigated to gain insight into the variation of the time-varying channel parameters, $\{b_{m,p,t}\}_{p \in \mathcal{P}}$.

9.3 RIR variation with changing source-sensor positions

To introduce time-variation in the all-pole channel model, a dynamic should be induced in the pole positions. To investigate the pole variation of an acoustic channel, the response of a $2.78 \times 4.68 \times 3.2$ office with reverberation time $T_{60} = 0.2$s is simulated at 16kHz sampling frequency using the ISM as discussed in sect. §4.3. One sensor is assumed at $\mathbf{r}_o = \begin{bmatrix} 1.6 & 1 & 1.3 \end{bmatrix}^T$ (i.e., at table height), whilst a 1.7m tall speaker moves

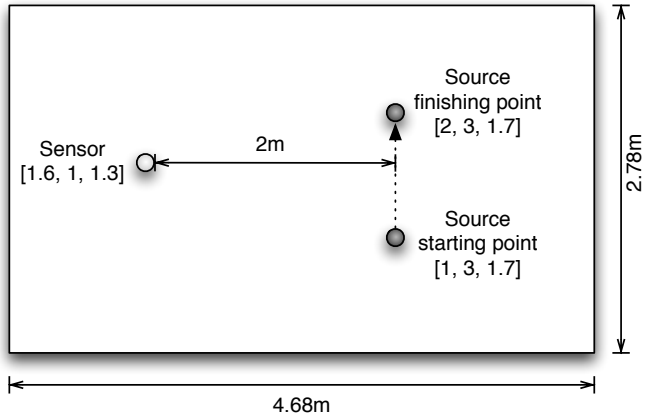


Figure 9.2: *Room setup*

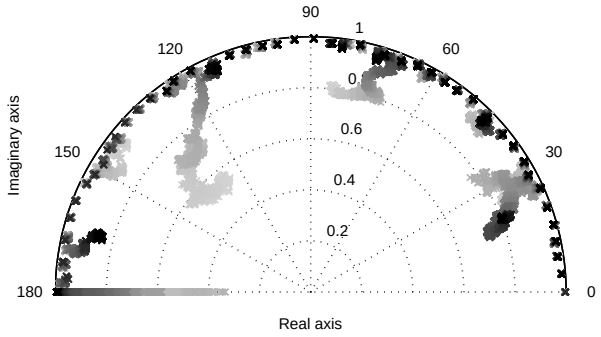
in a linear trajectory from $\mathbf{r}_{s,0} = \begin{bmatrix} 1 & 3 & 1.7 \end{bmatrix}^T$ to $\mathbf{r}_{s,T} = \begin{bmatrix} 2 & 3 & 1.7 \end{bmatrix}^T$ in 3.87ms (i.e., $T = 62,055$ samples at $f_s = 16$kHz) and using 100 steps (see Fig. 9.2 for an illustration of the room setup). The resulting audio signal can be found on the enclosed CD. In order to analyse the room response, the simulated RIR is modelled as an all-pole filter. In other words, the RIR is excited by WGN and the optimal model order of the response is extracted as described in sect. §4.4.2. The optimal model order is found as $P_{AIC} = 228$ using Akaike's information criterion (AIC) and $P_{MDL} = 100$ using the minimum description length (MDL). As discussed by, e.g., Wax and Kailath [219] or Liavas *et al.* [220], the AIC tends to over-model the channel order, whereas the MDL criterion is "shown to be asymptotically consistent" [220] and hence often favoured over the AIC.

An estimate of the all-pole filter coefficients, $\mathbf{b}_{0:t}$, is obtained by sliding a window of block size 5000 samples over the observed signal. In each window, the channel parameters are estimated by solving the Yule-Walker equations using the Levinson-Durbin recursion [221]. The corresponding pole trajectories are displayed in Fig. 9.3.

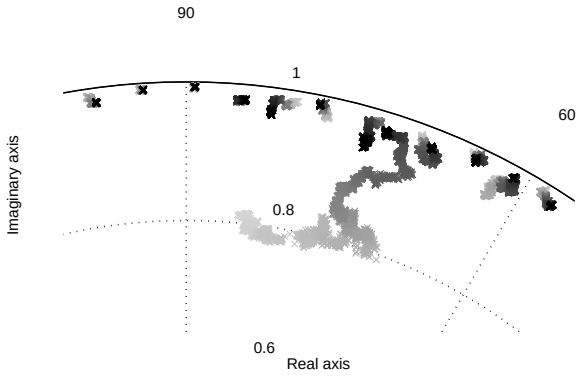
9.3.1 Channel pole variation with time

The variation of poles between $t = 5000, \ldots, 10,000$, plotted in Fig. 9.3a and magnified about 90 degrees phase in Fig. 9.3b, illustrates that the poles vary slowly and smoothly with time. It is thus desirable to incorporate the smooth and slow time variation of poles in the channel model.

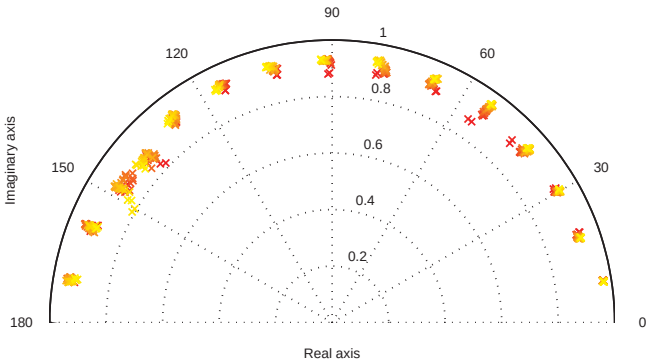
As part of the work in [1], the variation of the RIR with changing source-sensor positions was investigated by practical experiments. According to the layout in Fig. 9.4, the position of a sensor array consisting of 26 microphones was moved in 2mm inter-



(a) *Poles of the simulated RIR for Fig. 9.2. Time interval between* $\mathrm{t} = 5000, \ldots, 10000$*, showing variation of* $\mathrm{P}_{\mathrm{MDL}} = 100$ *poles.*



(b) *Zoom of poles in Fig. 9.3a, highlighting slow and smooth variation of poles.*



(c) *Poles of the measured RIR for Fig. 9.4. Red corresponds to early samples, yellow corresponds to samples towards the end of the sequence.*

Figure 9.3: *Pole variation extracted from simulated and measured RIR.*

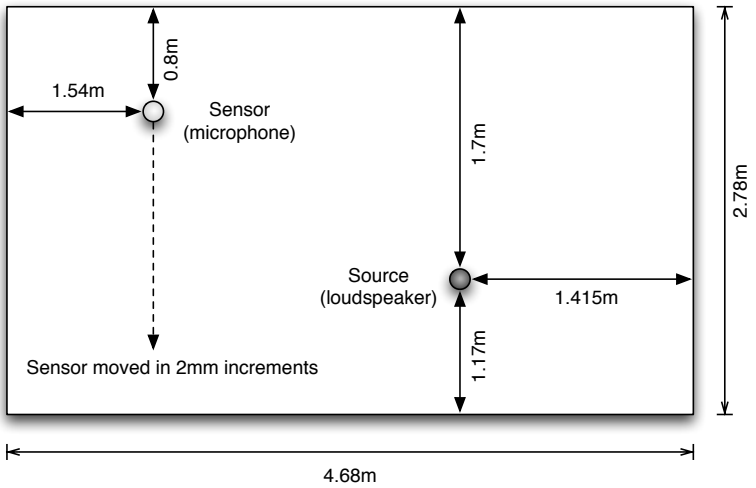


Figure 9.4: *Experimental setup for investigation of moving speakers as in [1].*

vals for a total change in distance of 7cm, whilst the source position was fixed. The impulse response of the room was measured for each change in the source-sensor position. As a result, a database of 32 impulse responses for each of the 26 microphones was created. For each increment, the response measured between the source and the seventh microphone is modelled as a subband all-pole filter between $0 - 4$kHz using the subband modelling approach in [222,223]. For each increment, the resulting model poles are plotted in Fig. 9.3c. The results illustrate that the poles of the RIR are located close to the unit circle and vary in a smooth and almost circular manner with small changes in position.

Similar to the argument in sect. §3.4.1 on page 49, the slow variation of the poles in Figures 9.3b and 9.4 could be modelled as by a first-order Markov chain with low variance on the pole positions. However, some poles exhibit significantly small variation and are almost static (see, e.g., the pole around 89 degress in Fig. 9.3b). An appropriately small variance on the random walk to reflect extremely small pole variations could lead to computational and numerical issues. Furthermore, in order to model the reverberant observed signal in eqn. (4.10), the AR parameters rather than the poles are required. As the relationship between poles and parameters is non-linear, a closed-form expression for the poles cannot be derived for high-order models. Therefore, it is desirable to model the parameters of the all-pole channel model directly, similar to the speech parameter models in sect. §3.4 on page 48.

The variation of the first, 51^{st} and 100^{th} parameter is shown in Fig. 9.5. Although the parameter trajectories exhibit small variations resembling WGN, the overall envelope of the trajectories varies relatively smoothly. The time-varying channel parame-

ters can be approximated by polynomial functions.

9.4 Polynomial approximation of the channel parameters

To investigate the accuracy of polynomial approximations of the channel coefficients, the curve fitting of the following polynomials to the parameter trajectories is performed:

- *Fourier polynomials:*

$$P_n^{\text{Fourier}} = b_n \cos(nt) + b_n \sin(nt) \tag{9.2a}$$

- Chebyshev polynomials: An orthogonal polynomial sequence related to de Moivre's formula and defined by the recursion:

$$P_n^{\text{Cheb}} = \begin{cases} 1 & \text{if } n = 0 \\ t & \text{if } n = 1 \\ 2t\, P_{n-1}^{\text{Cheb}} - P_{n-2}^{\text{Cheb}} & \text{if } n \geq 2 \end{cases} \tag{9.2b}$$

- *Legendre polynomials:* An orthogonal polynomial sequence forming the solution of the Legendre differential equation and defined by the recursion:

$$P_n^{\text{Legendre}} = \begin{cases} 1 & \text{if } n = 0 \\ t & \text{if } n = 1 \\ \frac{1}{n+1}\left[(2n+1)\, t\, P_{n-1}^{\text{Legendre}} - n P_{n-2}^{\text{Legendre}}\right] & \end{cases} \tag{9.2c}$$

- *Hermite polynomials:* An orthogonal polynomial sequence, used in statistics, physics,

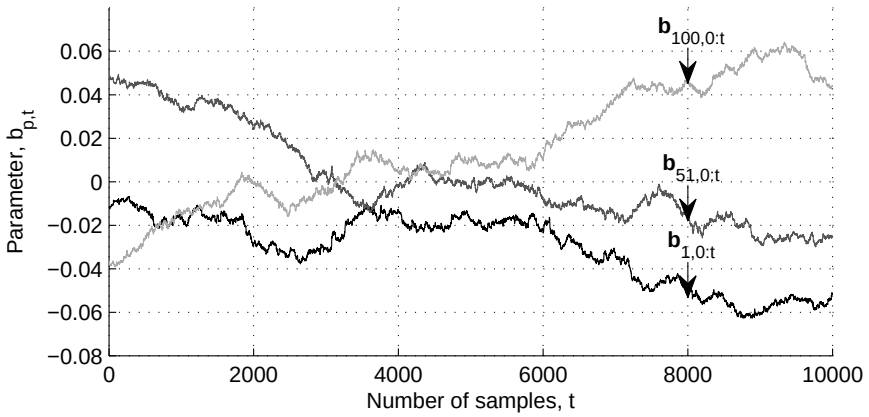


Figure 9.5: *Channel parameter variation with time, shown parameter* 1, 51, *and* 100.

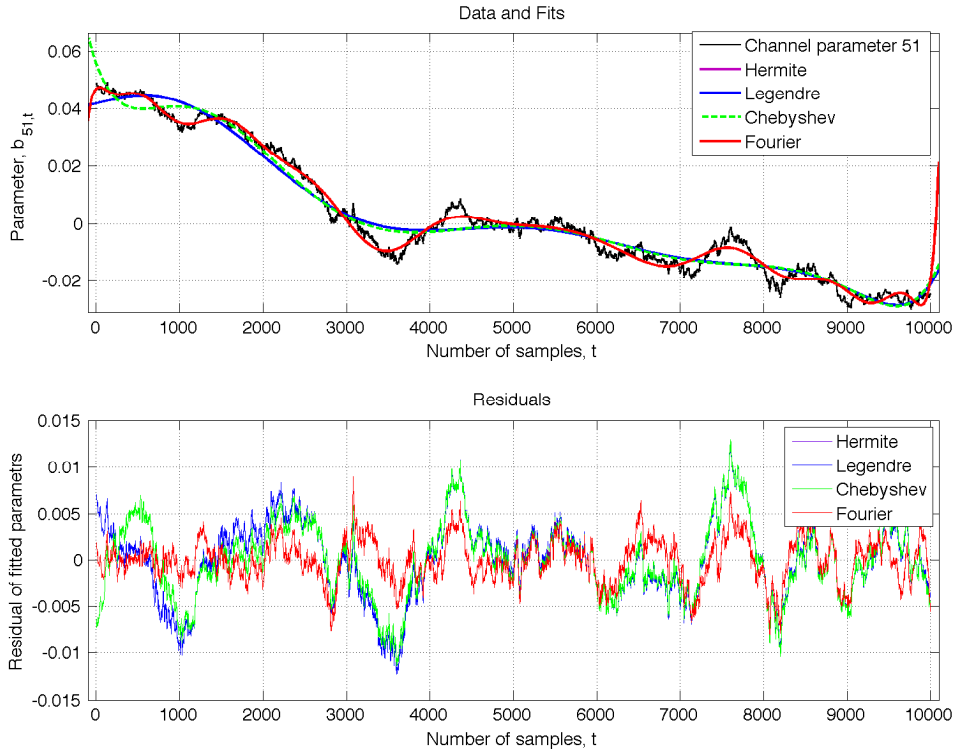


Figure 9.6: *Curve fitting of the* 51^{st} *channel parameter using Fourier, Chebyshev, Hermite, and Legendre polynomials. Approximation of the parameters is shown in the top graph, the residual of the fitted parameters is plotted in the bottom graph.*

and combinatorics.

$$P_n^{\text{Hermite}} = b_n(-1)^n e^{t^2/2} \frac{d}{dt^n} e^{-t^2/2}. \tag{9.2d}$$

Fig. 9.6 shows the fit of the 51^{st} parameter with the polynomial functions in eqn. (9.2a) to eqn. (9.2d) and the resulting residual, i.e., the difference between the data and its fit. The best fit is achieved using Fourier polynomials with a root mean squared error (RMSE) of 0.0024, closely approximating the general trend of the parameter variation. Chebyshev, Hermite, and Legendre polynomials cannot model sudden increases or decreases in the parameters. Note that the Hermite polynomials yield the same fit as Chebyshev polynomials. The partial misfit of the data is reflected in an RMSE of 0.004 for the Chebyshev and Hermite polynomial and an RMSE of 0.0042 for the Legendre polynomials.

Fitting the TVAR parameters by any of the polynomials in eqn. (9.2a) to eqn. (9.2d) is equivalent to projecting the time-varying coefficients, $\mathbf{b}_t$, to a space where they can

be considered as time-invariant coefficients, $\mathbf{b}$.

9.5 TVAP model by a linear combination of basis functions

Therefore, the TVAR parameters can be modelled as discussed in [128, 215–217] as

$$b_{p,t} = \sum_{\ell \in \mathcal{L}} b_{p,\ell}\, f_{t-p,\ell} = \mathbf{b}^T \mathbf{f}_{t-p} \tag{9.3}$$

where $\mathbf{b} = \{b_{p,\ell} : p \in \mathcal{P},\, \ell \in \mathcal{L}\}$ are the *unknown* time-invariant channel coefficients, and $\mathbf{f}_t \triangleq \begin{bmatrix} f_{t,1} & \dots & f_{t,L} \end{bmatrix}^T$ are the *known* time-varying basis functions.

The crux of eqn. (9.3) is that the basis functions, $\mathbf{f}_{t-p}$ span a space in which the time-varying parameters, $\{b_{m,p,t}\}$, can be considered as static parameters, $\{b_{m,p,\ell}\}$. The known basis functions therefore dictate the dynamic on the parameters. As the investigation in sect. §9.3 demonstrated, Fourier polynomials represent the time-varying channel parameters of a simulated RIR most accurately and are therefore used as basis functions for the representation of time-varying channels, i.e.,

$$\mathbf{f}_t = \begin{bmatrix} \sin(t) & \cos(t) & \dots & \sin(\frac{L}{2}t) & \cos(\frac{L}{2}t) \end{bmatrix}^T. \tag{9.4}$$

The observed signal in eqn. (9.1) can thus be formulated as

$$y_{m,t} = \sum_{p \in \mathcal{P}} \underbrace{\left\{ \sum_{\ell \in \mathcal{L}} b_{m,p,\ell}\, f_{\ell,t-p} \right\}}_{b_{m,p,t}} y_{m,t-p} + x_t + \sigma_{w_{m,t}}\, w_{m,t}, \tag{9.5}$$

This observation model can be easily rewritten in matrix form as

$$\mathbf{y}_t = \mathbf{Y}_{t-1}^T \mathbf{F}_t\, \mathbf{b} + \mathbf{C}^T \mathbf{x}_t + \mathbf{\Sigma}_{w_t}\, \mathbf{w}_t = \overline{\mathbf{Y}}_{t-1}^T \mathbf{b} + \mathbf{C}^T \mathbf{x}_t + \mathbf{\Sigma}_{w_t}\, \mathbf{w}_t \tag{9.8}$$

where $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}_{M\times 1}, \mathbf{I}_M)$, the channel parameters are defined as $\mathbf{b} \triangleq \begin{bmatrix} \mathbf{b}_1^T & \dots & \mathbf{b}_M^T \end{bmatrix}^T$ where $\mathbf{b}_m \triangleq \begin{bmatrix} b_{m,1,1} & \dots & b_{m,1,L} & \dots & b_{m,P,1} & \dots & b_{m,P,L} \end{bmatrix}^T$ are the coefficients corresponding to sensor $m \in \mathcal{M}$, the observation matrix is adjusted to incorporate the basis functions via $\overline{\mathbf{Y}}_{t-1}^T \triangleq \mathbf{Y}_{t-1}^T \mathbf{F}_t$, and $\mathbf{F}_t \triangleq \operatorname{diag}\begin{bmatrix} \mathbf{F}_{1,t} & \dots & \mathbf{F}_{M,t} \end{bmatrix}$ contains the basis func-

for $t > \max\{P, Q\}$ **do**
  for $i = 1, \ldots, N$ **do**
1 Importance sampling of $\theta_{0:t}^{(i)}$;
2 Kalman filter prediction of $\boldsymbol{\mu}_{t|t-1}^{(i)}$, $\boldsymbol{\Sigma}_{t|t-1}^{(i)}$ (eqns. (6.21a) and (6.21b)).;
3 Kalman filter estimation of $\boldsymbol{\mu}_{\mathbf{b},t}^{(i)}$ and $\boldsymbol{\Sigma}_{\mathbf{b},t}^{(i)}$ (eqn. (6.22)) using $\tilde{\mathbf{Y}}_{t-1}^{\star\,\mathsf{T}} \triangleq \overline{\mathbf{Y}}_{t-1} + \mathbf{C}^{\mathsf{T}}\boldsymbol{\Gamma}_{t|t-1}$:

$$\boldsymbol{\mu}_{\mathbf{b},t} = \left(\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_t}\tilde{\mathbf{Y}}_{t-1}^{\star\,\mathsf{T}}\right)\boldsymbol{\mu}_{\mathbf{b},t-1} + \mathbf{K}_{\mathbf{b}_t}\tilde{\mathbf{y}}_t \tag{9.6}$$

$$\boldsymbol{\Sigma}_{\mathbf{b},t} = \left(\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_t}\tilde{\mathbf{Y}}_{t-1}^{\star\,\mathsf{T}}\right)\boldsymbol{\Sigma}_{\mathbf{b},t-1}, \tag{9.7}$$

4 Kalman filter correction of $\boldsymbol{\mu}_{t|t}^{(i)}$, $\boldsymbol{\Sigma}_{t|t}^{(i)}$ (eqns. (6.21c) and (6.21d));
5 Evaluation of weights $w_t^{(i)}$ (eqns. (6.38), (6.32)):

$$p(\mathbf{y}_t \,|\, \mathbf{y}_{1:t-1}, \theta_t) = \mathcal{N}\left(\mathbf{y}_t \,\middle|\, \overline{\mathbf{Y}}_{t-1}\boldsymbol{\mu}_{\mathbf{b},t-1} + \mathbf{C}^{\mathsf{T}}\left(\boldsymbol{\alpha}_{t|t-1} + \boldsymbol{\Gamma}_{t|t-1}\boldsymbol{\mu}_{\mathbf{b},t-1}\right),\, \boldsymbol{\Sigma}_{\mathbf{z}_{t,\mathbf{b}}}\right)$$

  end
6 Normalization of importance weights;
7 Resampling;
8 Computation of particle average:

$$\widehat{\mathbf{x}}_t = \sum_{i\in\mathcal{N}} \hat{\boldsymbol{\mu}}_{t|t}^{(i)} \qquad \widehat{\theta}_t = \sum_{i\in\mathcal{N}} \theta_{0:t}^{(i)}, \qquad \widehat{\mathbf{b}} = \sum_{i\in\mathcal{N}} \boldsymbol{\mu}_{\mathbf{b},t}^{(i)}.$$

end

Algorithm 9.1: RBPF for moving speakers.

tions for each sensor, $m \in \mathcal{M}$, where

$$\mathbf{F}_{m,t} \triangleq \begin{bmatrix} f_{m,1,t-1} & \ldots & f_{m,G,t-1} & 0 & & & & \ldots & 0 & \\ 0 & \ldots & 0 & f_{m,1,t-2} & \ldots & f_{m,G,t-2} & 0 & \ldots & 0 & \\ \vdots & & \ddots & & \ldots & & & & \vdots & \\ 0 & & & & \ldots & & 0 & f_{m,1,t-P} & \ldots & f_{m,G,t-P} \end{bmatrix}$$

Comparing eqn. (9.8) to the observation space utilised in eqn. (6.9b) on page 115, i.e.,

$$\mathbf{y}_t = \mathbf{Y}_{t-1}\,\mathbf{b} + \mathbf{C}^{\mathsf{T}}\mathbf{x}_t + \boldsymbol{\Sigma}_{w_t}\mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}_{M\times 1}, \mathbf{I}_M), \tag{9.9}$$

the observation model employing the moving speaker in eqn. (9.8) is identical in form to eqn. (6.9b) with $\mathbf{Y}_{t-1}^{\mathsf{T}}$ redefined to $\overline{\mathbf{Y}}_{t-1}^{\mathsf{T}} = \mathbf{Y}_{t-1}^{\mathsf{T}}\,\mathbf{F}_t$. Therefore, the RBPF as derived in Chap. 6 can be straightforwardly applied by redefining the matrix $\mathbf{Y}_{t-1}$ according to eqn. (9.8).

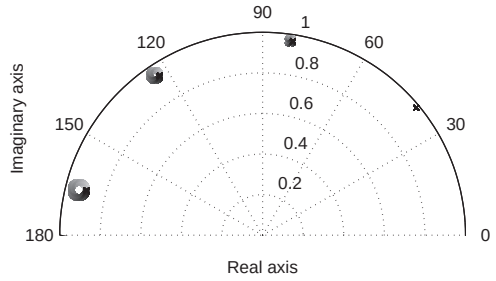


Figure 9.7: *Pole varying in a circular motion around the positions of the gramophone horn response.*

The RBPF applied to the blind dereverberation of moving speakers is therefore modified as outlined in Alg. 9.1. The resulting implementation of the RBPF is used in sect. §9.6 for experiments on blind speech dereverberation from time-varying channels.

9.6 Experimental results

The experiments in this section are based on the observation that small changes in the source-sensor position vary in a smooth and almost circular way for the measured response as shown by the pole trajectories in Fig. 9.3c. Therefore, a synthetic time-varying channel is generated whose poles revolve around the pole positions of the 8-th order gramophone horn response in a circular motion as illustrated in Fig. 9.7.

9.6.1 Experiments using synthetic source signals

A synthetic source signal of $Q = 15$ is generated according to the dynamic TVAR parameter model in Chap. 7. The signal is distorted by WGN with variance varying according to a random walk and of signal-to-noise ratio (SNR) 35dB. The resulting noisy signal is filtered with the channel in Fig. 9.7. The SRR of the observed signal filtered with the circular channel poles is -4.68dB or a LSD of 1.32dB.

The RBPF is executed using 2000 particles. The estimated channel poles and a comparison of the source signal with the observed signal and estimated signal are plotted in Figures 9.8 and 9.9a. The SRR of the estimated signal for the circular channel is -0.25dB and the LSD is 0.92dB yielding an SRR improvement 4.43dB and a LSD improvement of 0.4dB over the observed signal.

Although the channel poles do not mimic the pole variation precisely, the poles converge towards the centre of the circular poles in Fig. 9.7 and vary in a, to some ex-

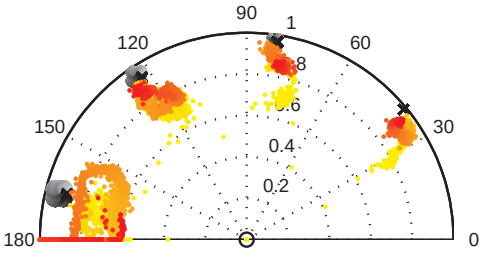


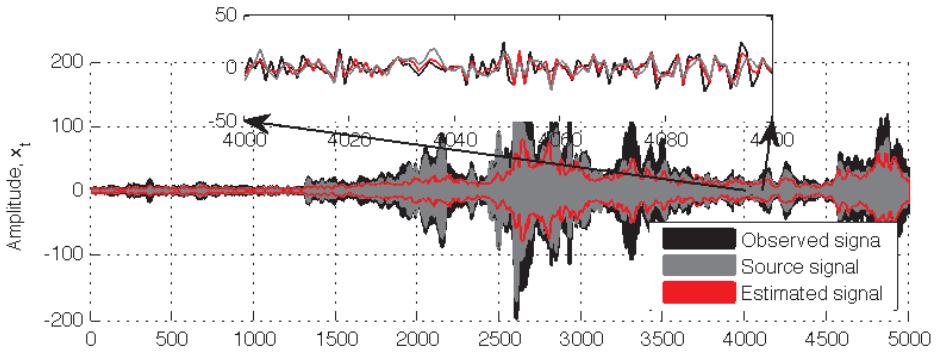
Figure 9.8: *Comparison of actual (gray) and estimated (coloured) channel pole trajectory for circular varying channel parameters for synthetic data generated with model order* $Q = 15$ *between* $t = P$ *(yellow) and* $t = 5000$ *(red).*

tent, circular motion around the centre points. Although the channel poles are therefore reasonably well captured, the envelope of the estimated signal approximates that of the source signal accurately only in certain segments, e.g., between $3200-3700$ samples or between $4500-5000$ samples (see Fig. 9.9a). To reiterate this point, Fig. 9.9b plots the instantaneous SRR,

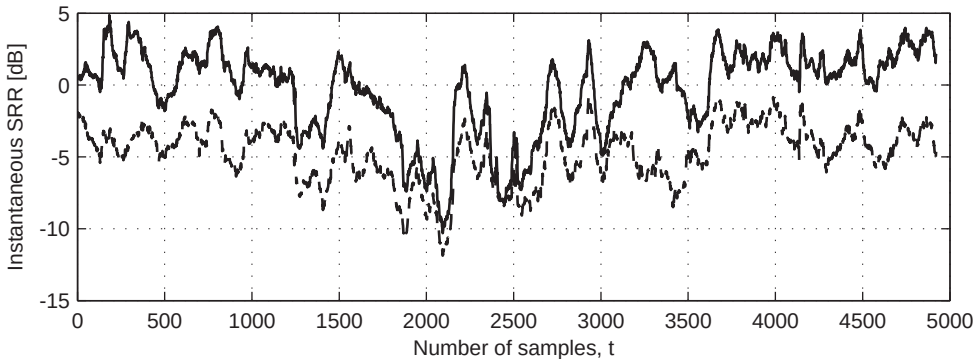
$$\text{iSRR}_{\text{dB}} \triangleq \frac{1}{M}\sum_{k=1}^{t} 10\,\log_{10}\left\{\frac{\sum_{\ell=Lk}^{Lk+L-1} x_\ell^2}{\sum_{\ell=Lk}^{Lk+L-1}(x_\ell-\chi_\ell)^2}\right\}. \tag{9.10}$$

that applies a sliding window of step size 1 over the signals and computes the SRR in each window. The instantaneous SRR in Fig. 9.9b highlights that that the estimate obtained using the RBPF approximates the source signal with SRRs up to 5dB (i.e., an improvement of 9.68dB as compared to the observed signal) for the first and the last 200 samples. Between 1500 and 3000 samples, the instantaneous SRR plummets up to values below 0dB, reaching up to -10dB. Therefore, even though some segments in the distorted observed signal can be enhanced effectively with significant SRR improvements, other segments cannot be recovered, resulting in the low overall segmental SRR value of -0.25dB.

It is desirable to investigate the reasons for the vast difference in enhancement performance of the RBPF for different segments in the signal. A possible reason could be that signal parameters vary slowly in sections with low resulting SRR such that slowly varying channel parameters cannot be identified from the source parameters as discussed in sect. §9.1. From Figures 9.9a and 9.9b, it is observed that the SRR decreases from 0dB to -8dB between $1500-2000$ samples and reaches values of up

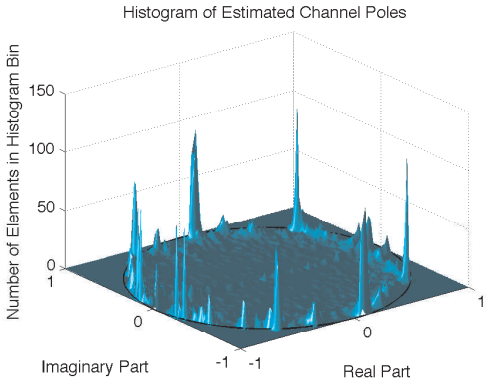


(a) *Comparison of the anechoic, reverberant, and estimated speech signal, "In the long run, it pays to buy quality clothing." at* $f_s = 4kHz$*, between* $0.2 - 0.5s$

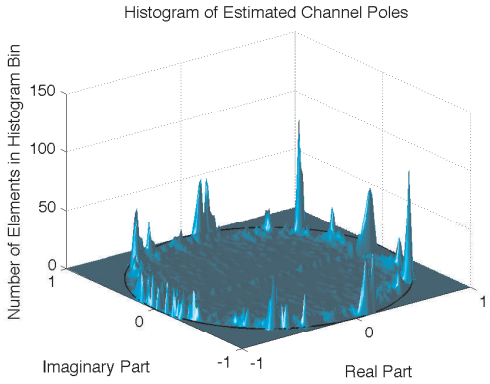


(b) *Instantaneous SRR of the estimated and observed signal*

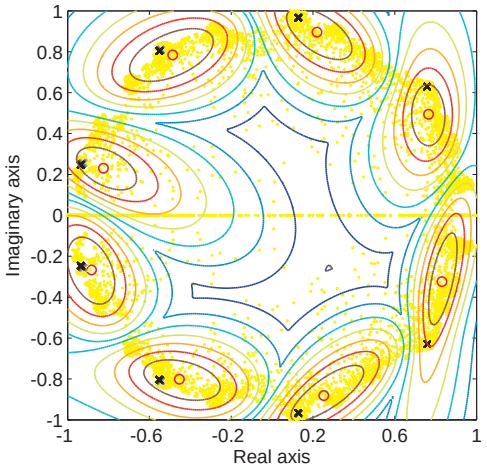
Figure 9.9: *Results for circular varying channel parameters for synthetic data generated with model order* $Q = 15$*.*



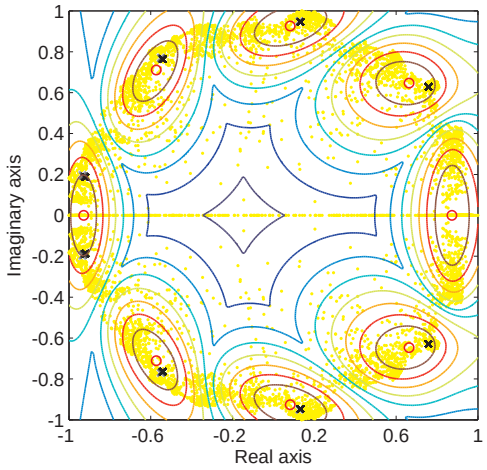
(a) *Histogram between* 1500 − 2000 *samples.*



(b) *Histogram between* 3000 − 3500 *samples.*



(c) *Clustering between* 1500 − 2000 *samples.*

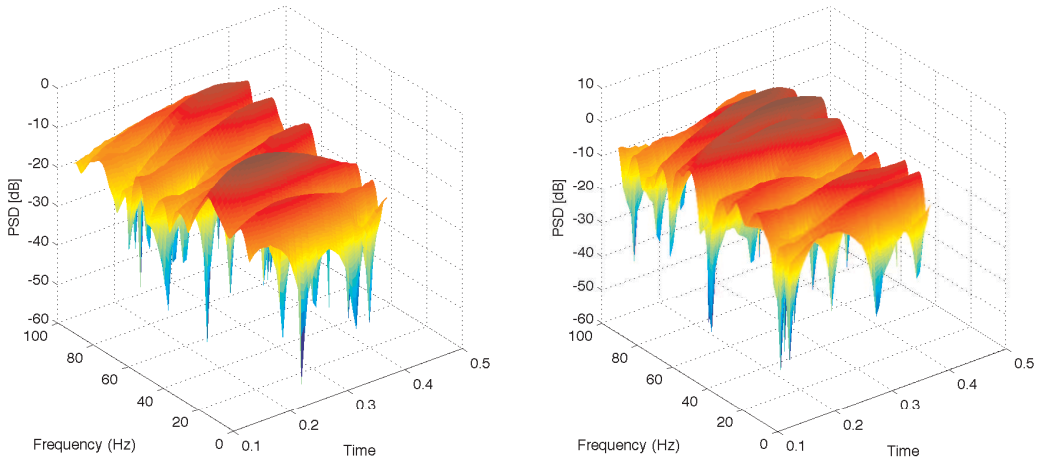


(d) *Clustering between* 3000 − 3500 *samples.*

Figure 9.10: *Yellow dots indicate histogram bins, red circles identify the peaks picked by the clustering algorithm, grey to black crosses indicate the trajectory of the actual channel poles in the corresponding segment. Coloured lines denote the contour plot of the clustered areas.*

to 5dB between 3000 − 3500 samples. The histogram method is hence applied to the observed signal in these regions to investigate whether the pole locations still form distinct peaks.

The resulting histograms are plotted in Figures 9.10a and 9.10b assuming $Q = 15$ source parameters and $P = 8$, i.e., extracting a total of 23 parameters. Both histograms indicate that distinct peaks close to the locations of the actual poles can be identified. To confirm this claim, k-mean clustering [224] was applied to the pole histogram, identifying the peak locations. The corresponding clusters are plotted in Figures 9.10c and 9.10d. In both signal segments, the channel trajectories are located closely to the peaks identified by the clustering algorithms apart from poles close to the real axis. There-



(a) *PSD of the source signal between* 1500 − 2000 *samples.*

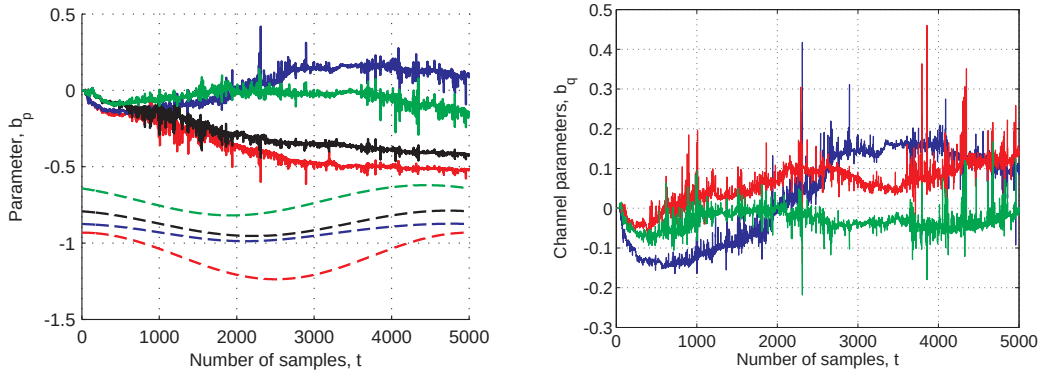
(b) *PSD of the source signal between* 3000 − 3500 *samples.*

Figure 9.11: *PSD of the source signal in different speech segments.*

fore, the channel model varies slowly, creating peaks in the histograms. In contrast, the source model varies rapidly such that the source poles smear over the floor of the histogram. The channel poles can thus be theoretically identified using the histogram method.

The result that the channel poles are, indeed, extractable from the observed signal using the histogram method suggests that the underlying problem does not lie with the channel poles. Instead, the source signal spectrum is examined. Therefore, Fig. 9.11 plots the power spectral density (PSD) of the source signal between 1500 − 2000 samples and between 3000 − 3500 samples. According to these graphs, the PSD of the source signal between 1500 − 2000 samples contains significantly less energy than the segment between 3000 − 3500 samples. In fact, the maximum of the PSD in Fig. 9.11a is −1.78dB, whereas the maximum of Fig. 9.11b is 5.13dB. I.e., the signal segment between 3000 − 3500 samples has a 6.91dB higher power content than the segment between 1500 − 2000 samples. Therefore, the power contained the segment between 3000 − 3500 samples is $10^{6.91/10} \approx 4.91$ times the power of the signal between 1500 − 2000 samples.

These results suggest that the dips in the trajectory of the instantaneous SRR, and hence the overall low segmental SRR, is due to weak power of the source signal in the corresponding signal segments. For weak signal power, the observed signal is severely dominated by the reverberant distortion and noise, such that recovery of the source signal can be difficult. Also, an issue with the SRR measure itself is encoun-



(a) *Comparison of actual (dashed) and estimated (solid) channel parameter trajectories,* $\mathrm{b}_{\mathrm{p},\mathrm{t}} = \mathrm{b}_{\mathrm{p},\ell}\, \mathrm{f}_{\ell,\mathrm{t}-\mathrm{p}}$*, showing* $\mathbf{b}_{1,0:\mathrm{t}}$ *(red),* $\mathbf{b}_{3,0:\mathrm{t}}$ *(blue),* $\mathbf{b}_{6,0:\mathrm{t}}$ *(black) and* $\mathbf{a}_{8,0:\mathrm{t}}$ *(green).*

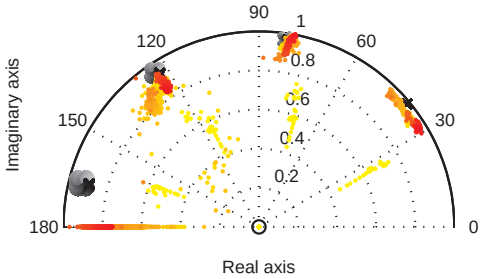
(b) *Static estimated channel parameters,* $\mathrm{b}_{\mathrm{p},\ell}$*, for* $\mathrm{p} = 3$ *(blue),* $\mathrm{p} = 20$ *(red), and* $\mathrm{p} = 40$ *(green).*

Figure 9.12: *Estimated channel parameters*

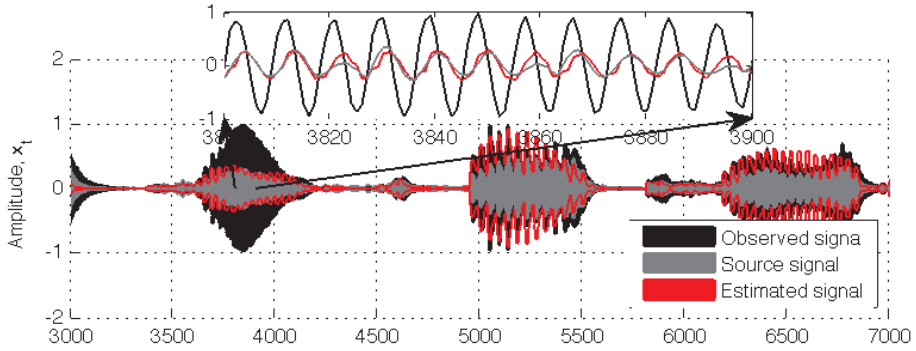
tered: the segmental SRR in intervals of small signal energy takes on large negative values, thus biasing the overall measure [123]. In order to circumvent this effect, silent frames should either be excluded from the segmental SRR, e.g., using voice activity detector (VAD) [210, 225].

Although weak signal powers were also encountered for the static channel, blind dereverberation in this case is based on a channel estimate that after a limited time duration converge towards a steady state. If segments of low signal power are encountered after the convergence of the channel, source signal estimation is more robust and hence less prone to the deleterious effects of the dominating distortion in the segment. However, for the moving speaker where the channel estimates are continuously changed, the signal estimator is less robust and cannot counter the dominating distortion.

It might be argued that according to the basis function model of the channel parameters, the time-varying channel parameters are projected to a space where they can be considered as a linear combination of *static* unknown parameters with time-varying known basis functions and that, therefore, also the time-varying channel parameters in this model should reach convergence in this model. However, exactly for the reason that the basis function representation of time-varying channel parameters is a *model*, the convergence of the underlying static parameters cannot be guaranteed if generation of the actual underlying channel parameters used for signal distortion does not adhere to the basis function model. This point can be clarified by plotting the estimated channel parameters. Fig. 9.12 therefore plots the estimated time-varying chan-



(a) *Comparison of actual and estimated channel trajectory*



(b) *Comparison of the anechoic, reverberant, and estimated speech signal, "In the long run, it pays to buy quality clothing." at* $f_s = 4kHz$*, between* $0.2 - 0.5s$

Figure 9.13: *Results for circular varying channel parameters*

nel parameters, $b_{p,t} = b_{p,\ell}\, f_{\ell,t-p}$ (Fig. 9.12a) and the *static* parameters, $b_{p,\ell}$ (Fig. 9.12b). As can be seen from Fig. 9.12b, the parameters, $b_{p,\ell}$ do not converge to a constant value, but exhibit significant variation in amplitude change of up to 0.4 instead.

9.6.2 Experiments using speech data

The experiment in sect. §9.6.1 is repeated for the speech signal "In the long, it pays to buy quality clothing." at 4kHz. The SRR of the observed signal is -5.55dB.

The RBPF in Alg. 9.1 on page 182 is evaluated for 2000 particles using the dynamic TVAR parameter model in Chap. 7 assuming a speech model order of $Q = 15$. The estimated SRR is -2.97dB, an improvement of 2.57dB. The estimated channel parameter trajectory is displayed in Fig. 9.13a. Again, although the channel poles do not mimic the exact dynamic variation of the poles in Fig. 9.7, they converge towards the circle centres and vary closely around the area of the actual, underlying pole positions. The estimated signal encounters similar issues as in sect. §9.6.1, whereby the underlying anechoic speech signal is partially well recovered, but remains buried in the distortion by the channel and noise in a considerable proportion of segments. An enlarged ver-

sion of an accurately modelled section is displayed in Fig. 9.13b, illustrating that the envelope of the estimated signal approximates the envelope of the source signal. The zoom indicates that the detailed structure of the speech signal is captured accurately by the estimated signal.

Audio samples of the anechoic speech signal, the distorted observed signal, and the estimated signal can be found on the attached CD in the folder '*Chapter 7 - Moving speaker*'. The distorted signal exhibits a distant, metallic sound. The estimated signal sounds less distant and partially removes the metallic sound particularly at the beginning and towards the end of the signal. However, segments in the middle of the signal remain buried in the metallic sound of the observed signal.

9.7 Discussion

The RIR was also shown to be dependent on both the source and sensor position and therefore varies with changing source-sensor distances. Hence, for moving speakers, the RIR varies with the change in speaker location. As the speaker location varies with time, the RIR varies with time. It was investigated how the time-varying properties of the RIR for moving speakers is reflected in the poles and corresponding parameters of the all-pole channel filter. It was found that both the poles and parameters vary smoothly and slowly with time. A model based on a linear combination of basis functions was proposed for capturing the slow variation of the channel parameters. The accuracy for modelling the time-varying channel parameters of several polynomial functions was tested. Fourier functions were found to be most adequate in capturing sudden changes in the coefficients. These results are utilised in Chap. 9, where the blind dereverberation approach for stationary speakers is extended to moving speakers.

This chapter introduced an extension of the channel model used for the RBPF to facilitate blind dereverberation of speech from moving speakers. Measured and simulated experiments of moving speakers were used to demonstrate that the RIR changes significantly with changes in the source-sensor position. As the source-sensor positions change with time, the RIR changes with time, implying that the channel parameters characterising the RIR are time-varying as well. Therefore, as a speaker moves throughout a room, the RIR characterising the channel between the initial speaker location will be sufficiently different the RIR characterising the speaker's location after a few steps. Therefore, the assumption of a static channel model is not sufficient to accurately model the reverberant RIR. The channel should therefore be modelled as time-varying instead.

It was shown in this chapter that the time-varying parameters of an all-pole model of a RIR between a sensor and moving source can be approximated by Fourier polynomials. Therefore, it was proposed to model the time-varying channel parameters as a combination of static unknown parameters with a time-varying known set of Fourier basis functions. Due to the general form of the RBPF, the model can be easily implemented in the proposed speech dereverberation framework.

Using the resulting RBPF implementation, experimental results based on synthetic and speech data were conducted. Results indicated that the trajectory of channel poles is estimated reasonably well. The anechoic source signal is approximated accurately by the estimated signal in signal segments where the signal power is sufficient. However, in regions of small signal power, the reverberant distortion cannot be removed.

The issue encountered with low signal powers can be circumvented by using a variant of VAD, whereby only sections with sufficient signal power are processed [225]. In periods of low source signal power or silence periods, the RBPF stops to update the unknown variables, $\boldsymbol{\varphi}_{0:t}$ and "remembers" the states of the last signal segment of sufficiently large signal power.

Chapter 10

Computational complexity and extension to multirate processing

10.1 Introduction

Whilst particle filters are well suited for state estimation of non-linear state-spaces due to their sample-based representation, the improved state space representation comes at the cost of increased computational complexity. As shown in this chapter, whilst the computational complexity of the Rao-Blackwellized particle filter (RBPF) is linear in the number of particles and samples, complexity increases quadratically in the number of channel parameters and number of sensors. Realistic room impulse responses (RIRs) require several hundred autoregressive (AR) coefficients for accurate representation of the response by an all-pole filter. Therefore, the computational overhead of the RBPF particle in its vanilla form as discussed in Chaps. 6 and 8 becomes severe when applied to speech distorted in realistic rooms, in particular when utilising microphone arrays.

The question thus arises how to address real-time or near real-time applications where the refresh rate of new incoming data is higher than the update rate of the particle filter. It was noted in the tracking community that adaptive sampling can be applied to improve the computational efficiency of particle filters. A large number of samples is required initially to estimate the target position. Nonetheless, as soon as an accurate estimate of the target location is available, only few samples are necessary to actually track the position with time. As the accuracy of the location estimates improves, the likelihood of the importance weights increases. Therefore, if the importance weights are of high value, the positional estimates are well in tune with the actual target position and fewer samples are required. The number of samples can

thus be reduced if the sum of weights exceeds a certain threshold [226, 227].

Adaptive sampling is an interesting approach for the reduction of computational complexity in state spaces that might contain sparsity, such as localisation and target tracking applications. However, speech enhancement generally involves non-sparse state spaces, such that adaptive sampling would lead to a degradation in the quality of the reconstructed signal due to, essentially, an adaptively altered sampling frequency.

Alternatively, the marginalisation of the channel and model parameters from the joint probability density function (pdf) to obtain the marginal pdf of the source signal,

$$p\left(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right) = \int_{\mathbb{R}^Q} \int_{\mathbb{R}^P} p\left(\mathbf{x}_{0:t}, \boldsymbol{\theta}_{0:t}, \mathbf{b} \mid \mathbf{y}_{1:t}\right) \mathrm{d}\mathbf{b}\, \mathrm{d}\boldsymbol{\theta}_{0:t} \tag{10.1}$$

can be approximated using a variational Bayes (VB) approach [228]. In the VB approach, $p\left(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right)$ is obtained by approximating the joint posterior density by

$$p\left(\mathbf{x}_{0:t}, \boldsymbol{\theta}_{0:t}, \mathbf{b} \mid \mathbf{y}_{1:t}\right) \approx p\left(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right) p\left(\boldsymbol{\theta}_{0:t} \mid \mathbf{y}_{1:t}\right) p\left(\mathbf{b} \mid \mathbf{y}_{1:t}\right) \tag{10.2}$$

and minimising the Kullback-Leibler distance of both sides. Using principal component analysis (PCA) and singular value decomposition (SVD) approaches, computational efficiency can be improved. Nonetheless, VB approaches only approximate the solution already derived in Chaps. 6 and 8 and thus lead inferior estimates.

An approach for the reduction in computational complexity applicable to a wide range of audio processing algorithm and not limited to Bayesian inference, is multirate filtering of the received signal [229–238]. The observed signal is divided into several frequency subbands using an analysis filterbank. Dereverberation can be performed for each subband signal. The final estimated signal is then reconstructed using a synthesis filterbank that recombines the enhanced subband signals to the fullband signal. Subband filtering was effectively applied in the literature for dereverberation problems in, e.g., [36, 56, 80, 213, 222, 223].

This chapter adapts subband filtering for the RBPF in order to reduce the computational complexity. Sect. §10.2 derives the computational complexity of the RBPF as proposed in Chap. 6. Sect. §10.4 investigates the channel order required for realistic RIRs and shows that the results lead to severe computational burden of the RBPF. Results are discussed in sect. §10.5.

10.2 Computational complexity

The results in Chap. 7 to Chap. 9, required several hundred particles for accurately approximating the multidimensional state space in eqn. (6.9) on page 115. It is therefore of interest to evaluate the computational complexity of the RBPF in order to evaluate its computational efficiency or burden.

For the evaluation of the computational complexity of the RBPF, the dynamic time-varying AR (TVAR) parameter and stationary channel model in Chap. 7 are assumed for brevity of this chapter. Nonetheless, the results are easily extendible to e.g., the parallel formant synthesizer (PFS) source model or the moving speaker. Based on Alg. 7.1 on page 130, the following equations should therefore be estimated for each time step, $t \geq 1$, and particle, $i \in \mathcal{N}$:

- **Importance sampling of the source parameters and covariance terms:**

$$\mathbf{a}_t^{(i)} = \mathbf{a}_{t-1}^{(i)} + \boldsymbol{\Sigma}_{\mathbf{a}_t}\, \mathbf{r}_{\mathbf{a}_t} \tag{10.3}$$

$$\phi_{\nu_t}^{(i)} = \phi_{\nu_{t-1}}^{(i)} + \sigma_{\phi_{\nu_t}}\, r_{\phi_{\nu_t}} \tag{10.4}$$

$$\boldsymbol{\phi}_{\mathbf{w}_t}^{(i)} = \boldsymbol{\phi}_{\mathbf{w}_{t-1}}^{(i)} + \boldsymbol{\Sigma}_{\boldsymbol{\phi}_{\mathbf{w}_t}}\, \mathbf{r}_{\boldsymbol{\phi}_{\mathbf{w}_t}} \tag{10.5}$$

- **Prediction of the Kalman filter equations:** According to eqns. (6.21a) and (6.21b) on page 119:

$$\boldsymbol{\mu}_{\mathbf{x}_{t|t-1}} = \mathbf{A}_t\, \boldsymbol{\mu}_{\mathbf{x}_{t-1|t-1}} \tag{10.6}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}} = \mathbf{A}_t^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{x}_{t-1|t-1}} \mathbf{A}_t + \boldsymbol{\Sigma}_{\mathbf{v}_t}\, \boldsymbol{\Sigma}_{\mathbf{v}_t}{}^{\mathsf{T}} \tag{10.7}$$

$$\mathbf{K}_{\mathbf{x}_t} \triangleq \left(\mathbf{A}_t^{\mathsf{T}}\, \boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{t-1}}\, \mathbf{Y}_{t-1}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}}\, \mathbf{C}\right) \boldsymbol{\Sigma}_{z_t}^{-1} \tag{10.8}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{z_t} = {} & \mathbf{Y}_{t-1}\, \boldsymbol{\Sigma}_{\mathbf{b},t-1}\, \mathbf{Y}_{t-1}^{\mathsf{T}} + \mathbf{C}^{\mathsf{T}}\, \boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}}\, \mathbf{C} + \boldsymbol{\Sigma}_{w_t}\, \boldsymbol{\Sigma}_{w_t}^{\mathsf{T}} \\ & + \mathbf{C}^{\mathsf{T}}\, \mathbf{A}_t^{\mathsf{T}}\, \boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{t-1}}\, \mathbf{Y}_{t-1}^{\mathsf{T}} + \mathbf{Y}_{t-1}\, \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{t-1}}\, \mathbf{A}_t\, \mathbf{C} \end{aligned} \tag{10.9}$$

- **Estimation of the channel parameters:** According to eqns. (6.22a) and (6.22b) on page 119:

$$\boldsymbol{\mu}_{\mathbf{b},t} = (\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_t}\, \mathbf{Y}_{t-1})\, \boldsymbol{\mu}_{\mathbf{b},t-1} + \mathbf{K}_{\mathbf{b}_t} \left(\mathbf{y}_{t-1} - \mathbf{C}^{\mathsf{T}}\, \boldsymbol{\mu}_{\mathbf{x}_{t|t-1}}\right) \tag{10.10}$$

$$\boldsymbol{\Sigma}_{\mathbf{b},t} = (\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_t}\, \mathbf{Y}_{t-1})\, \boldsymbol{\Sigma}_{\mathbf{b},t-1} - \mathbf{K}_{\mathbf{b}_t}\, \mathbf{C}^{\mathsf{T}}\, \mathbf{A}_t^{\mathsf{T}}\, \boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{t-1}} \tag{10.11}$$

$$\mathbf{K}_{\mathbf{b}_t} \triangleq \left(\boldsymbol{\Sigma}_{\mathbf{b},t-1}\, \mathbf{Y}_{t-1}^{\mathsf{T}} + \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{t-1}}\, \mathbf{A}_t\, \mathbf{C}\right) \boldsymbol{\Sigma}_{z_t}^{-1} \tag{10.12}$$

$$\tag{10.13}$$

- **Estimation of the marginalised mean and covariance:** According to eqns. (6.21c)

and (6.21d):

$$\boldsymbol{\mu}_{\mathbf{x}_{t|t}} = \left(\mathbf{I}_Q - \mathbf{K}_{\mathbf{x}_{t-1}} \mathbf{C}^T\right) \boldsymbol{\mu}_{\mathbf{x}_{t|t-1}} + \mathbf{K}_{\mathbf{x}_{t-1}} \left(\mathbf{y}_{t-1} - \mathbf{Y}_{t-1} \boldsymbol{\mu}_{\mathbf{b},t-1}\right) \tag{10.14}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}_{t|t}} = \left(\mathbf{I}_Q - \mathbf{K}_{\mathbf{x}_t} \mathbf{C}^T\right) \boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}} - \mathbf{K}_{\mathbf{x}_t} \mathbf{Y}_{t-1} \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{t-1}} \mathbf{A}_t \tag{10.15}$$

- **Estimation of the cross-correlation terms:** According to eqns. (6.26) and (6.27):

$$\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_t} = \left(\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_t} \mathbf{Y}_{t-1}\right) \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{t-1}} \mathbf{A}_t - \mathbf{K}_{\mathbf{b}_t} \mathbf{C}^T \boldsymbol{\Sigma}_{\mathbf{x}_{t|t-1}} \tag{10.16}$$

$$\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_t} = \left(\mathbf{I}_Q - \mathbf{K}_{\mathbf{x}_t} \mathbf{C}^T\right) \mathbf{A}_t^T \boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{t-1}} - \mathbf{K}_{\mathbf{x}_t} \mathbf{Y}_{t-1} \boldsymbol{\Sigma}_{\mathbf{b},t-1} \tag{10.17}$$

The corresponding number of additions, multiplications, inverses, and determinants are summarised in Table 10.1.

Note that these results do not include the normalisation of the weights, resampling or computation of the particle average. However, as normalisation of the weights is a scalar operation, the computational complexity compared to Table 10.1 is negligible. Averaging of the particles involves NQ additions and Q multiplications for $\hat{\mathbf{x}}_t$, $N(Q + M + 1)$ additions and $Q + M + 1$ multiplications for $\hat{\boldsymbol{\theta}}_t$, as well as NMP additions and MP multiplications for $\hat{\mathbf{b}}$. Again, compared to Table 10.1, the linear growths in N, Q, M and P required for particle averaging is negligible. The computational complexity of resampling algorithms is discussed by Bolic *et al.* in [239]: According to these results, the complexity of systematic, residual, residual-systematic, and partial resampling is linear in the number of particles, i.e., $\mathcal{O}(N)$. Assignment of the resampled particles according to the indices obtained by the resampling algorithms reduces to a sorting problem according to the specified N resampled indices. Sorting algorithms take $\mathcal{O}(N^2)$ in the worst case (e.g., insertion sort, bubble sort, or quick sort) and $\mathcal{O}(N \log N)$ in the best case (e.g., merge sort, or heap sort) [240]. Therefore, neither normalisation, resampling, or averaging of the particles contributes significantly to the rate of growth of the algorithm and are negligible compared to the operations in Table 10.1.

According to the results in Table 10.1, most computational complexity is involved in computing the channel covariance and gain, the source covariance, and the cross-correlation terms. As the exact growth depends on the values of the channel order, P, source order, Q, and number of microphones, M, the overall rate of growth is:

$$C = \mathcal{O}\left(\max\left\{Q(MP)^2,\, Q^2MP,\, M^3P,\, MP^3\right\}\right). \tag{10.18}$$

As mentioned in Chap. 7 to Chap. 9, the source order is generally assumed as $Q = 15$.

Term	Multiplications	Inv.	Rate of growth
$\mathbf{a}_t$	Q	0	$\mathcal{O}(Q)$
ϕ_{ν_t}	1	0	$\mathcal{O}(1)$
$\boldsymbol{\phi}_{\mathbf{w}_t}$	M	0	$\mathcal{O}(M)$
$\boldsymbol{\mu}_{t\|t-1}$	Q	0	$\mathcal{O}(Q)$
$\boldsymbol{\Sigma}_{t\|t-1}$	$2Q^2+1$	0	$\mathcal{O}(Q^2)$
$\mathbf{K}_{\mathbf{x}_t}$	$QPM+2QM^2$	1	$\mathcal{O}(\max\{QPM, QM^2\})$
$\boldsymbol{\Sigma}_{z_t}$	$(MP)^2+2M^2P+M+MP$	0	$\mathcal{O}(M^2P^2)$
$\boldsymbol{\mu}_{\mathbf{b},t}$	$2M^2P+(MP)^2+MP^2$	0	$\mathcal{O}(M^2P^2)$
$\boldsymbol{\Sigma}_{\mathbf{b},t}$	$MP^2+(MP)^3+MP+2(MP)^2$	0	$\mathcal{O}(MP^3)$
$\mathbf{K}_{\mathbf{b}_t}$	$M^2P^2+QMP+M^3P$	1	$\mathcal{O}(\max\{M^2P^2, M^3P\})$
$\boldsymbol{\mu}_{\mathbf{x}_{t\|t}}$	$Q+2QM+MP$	0	$\mathcal{O}(\max\{QM, MP\})$
$\boldsymbol{\Sigma}_{\mathbf{x}_{t\|t}}$	$2Q^2+Q^2MP+QP$	0	$\mathcal{O}(Q^2MP)$
$\boldsymbol{\Sigma}_{(\mathbf{b}\|\mathbf{x})_t}$	$2QMP+Q(MP)^2+MP^2$	0	$\mathcal{O}(Q(MP)^2)$
$\boldsymbol{\Sigma}_{(\mathbf{x}\|\mathbf{b})_t}$	$QP+Q(MP)^2+Q^2+Q^2MP$	0	$\mathcal{O}(\max\{Q(MP)^2, Q^2MP\})$

Table 10.1: *Operations required for computation of the RBPF sorted by sequence of execution according to Alg. 6.1.*

Furthermore, as will be shown in sect. §10.4, the channel order required to model RIRs lies between approximately $200-1000$ parameters. Therefore, the source order can be assumed to be significantly lower than the channel order for RIRs, i.e., $Q \ll P$. Thus, for full-band applications using relatively small sensor arrays of $M < 100$, $Q^2MP < M^3P < Q(MP)^2 < MP^3$.

Executing the RBPF for N particles and T time samples, the operations in Table 10.1, the rate of growth becomes

$$C_{N>1,t>1} = \mathcal{O}\left(T\,N\max\left\{Q(MP)^2,\ Q^2MP,\ M^3P,\ MP^3\right\}\right). \tag{10.19}$$

In order to confirm the quadratic / cubic growth in the number of sensors, consider the following experiment: $M = 1,\ldots,10$ synthetic AR channels of channel order $P = 10$ are generated by sampling $P/2$ points in the z-plane with radius 0.95 and randomly sampled phases within $[0,\pi]$. The corresponding channel poles are constructed by computing the $P/2$ sampled poles $p = r\exp j\omega$ and appending their $P/2$ complex conjugates to the pole vector. A second-order TVAR signal of 5000 samples is filtered with the all-pole filter to generate the observed signal. The RBPF is evaluated for $N = 150$ particles for $M = 1,\ldots,10$ sensors. The run time of the algorithm for the increasing number of channel orders is plotted in Fig. 10.1. The run time increases from approximately 580s for one sensor to 1100s for 10 sensors. The trend of the data is fit to a quadratic and cubic curve as indicated in the figure. Both the quadratic and cubic curve fit the data well, thus verifying the quadratic / cubic growth of complexity with

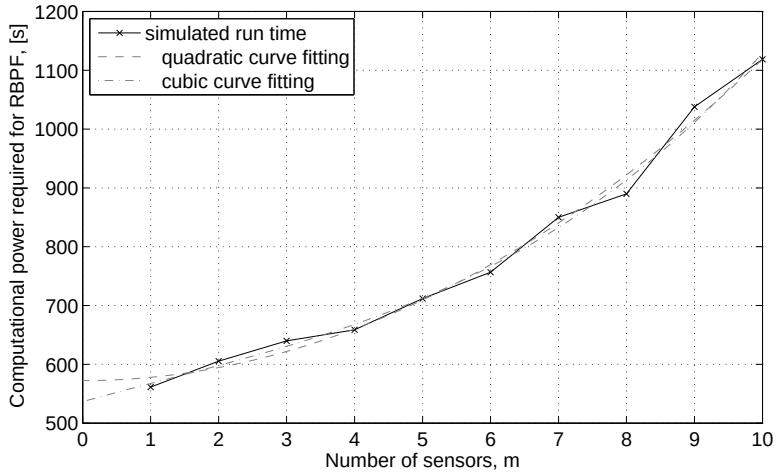


Figure 10.1: *Increasing run time with number of sensors and its quadratic and cubic approximation.*

the number of microphones described in eqn. (10.19).

Ideally, the quadratic growth in the channel order should be confirmed using a similar experiments where P increases from 2 to 1000.[1] MATLAB's `poly` function is required to transform the poles into AR parameters in order to filter the source signal with the channel response. Unfortunately, for model orders of $P > 100$ and poles located close to the unit circle, i.e., $r > 0.9$, `poly` causes numerical issues due to the non-linear transformation between the poles and parameters.

Based on the result in eqn. (10.19), the increase in computational complexity in the number of particles, sensors, and channel order should be discussed. Sect. §10.3 therefore discusses the rate of growth with the number of sensors, whilst sect. §10.4 discusses the rate of growth with the channel order for realistic RIRs.

10.3 Rate of growth vs. the number of sensors and particles

Assuming $Q = 15$ source parameters and $30,000$ samples (equivalent to 3.75s of speech at a sampling frequency of $f_s = 8$kHz), eqn. (10.19) is evaluated increasing number of particles, N and increasing number of microphones. The resulting growth rates are plotted in Fig. 10.2.

Due to the quadratic / cubic growth of C_{MARBLE} with M in eqn. (10.19), Fig. 10.2

[1] As will be shown in sect. §10.4, the channel order of realistic RIRs can take values of up to $P = 1000$ at 16kHz sampling frequency.

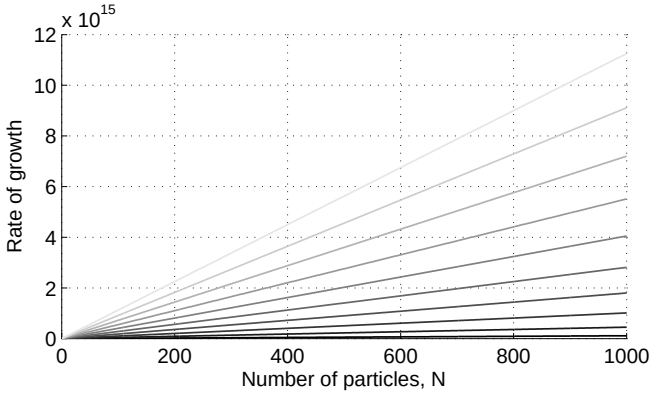


Figure 10.2: *Linear increase of the computational complexity with the number of particles,* N*, in eqn.* (10.19) *with the number of particles for a single sensor (—) up to* 10 *microphones (—) using a source order of* $Q = 15$ *and* $T = 30,000$ *samples.*

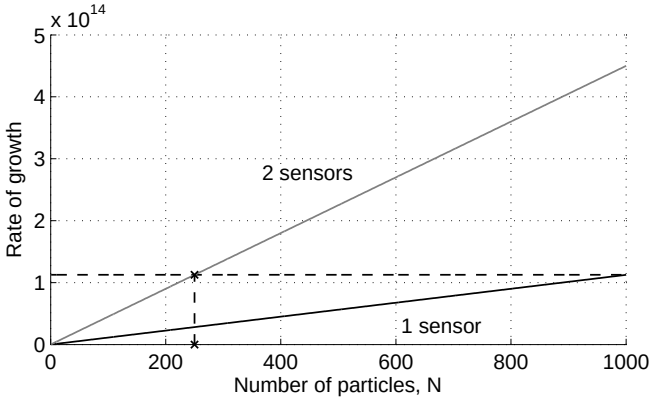


Figure 10.3: *Comparison between complexity of single and dual-sensor RBPF.*

shows that single-microphone processing is significantly less computationally demanding than multi-sensor processing. As multiple microphones utilise more statistical knowledge of the same event, it seems reasonable to argue that less particles are necessary to obtain the same accuracy of performance of single-sensor dereverberation. However, Fig. 10.2 reveals that regardless of the reduction of necessary particles due to increasing number of microphones, N, the computational complexity of single-sensor processing is significantly less. For instance, when using $M = 5$ sensors, as few as $N = 20$ particles would have to suffice for estimation with the same complexity order as a single sensor using $N = 1000$ particles.

As an example of the necessary reduction in the number of particles, the results in Fig. 10.2 are compared for a single and two microphones in Fig. 10.3. As an example assume an experiment using a single sensor requires 1000 particles for accurate source signal estimation. As shown in sect. §7.4, it was shown that the dereverberation perfor-

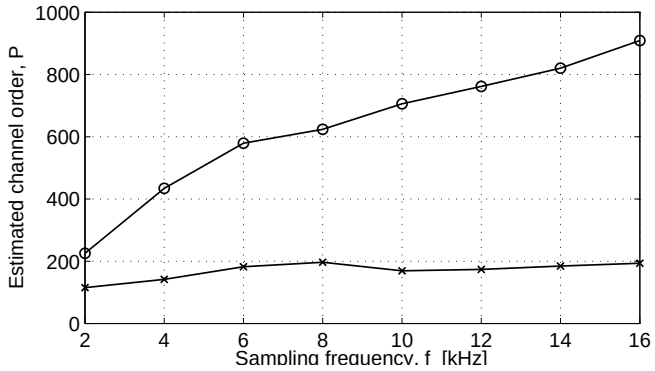


Figure 10.4: *Optimal channel order increasing with sampling frequency of simulated ISM room acoustics.*

mance of the RBPF can be improved by using multiple microphones. Therefore, consider introducing a second sensor to the experiment. However, the markers in Fig. 10.3 indicate that 250 particles should be used at most in order to obtain the same computational complexity as the single sensor case using 1000 particles. In other words, only if the additional knowledge inferred from a second sensor is sufficient to achieve accurate estimation using 250 particles as opposed to 1000 particles for a single sensor, multiple sensors are computationally as feasible as a single sensor. Therefore, a trade-off between improved dereverberation performance and computational burden has to be made.

According to eqn. (10.19), the computational complexity of the RBPF increases quadratically with the number of channel parameters, P. To elaborate on these results and to gain insight into the channel model orders required for accurate modelling of realistic RIRs, the following section investigates the optimal channel order required for simulated RIRs.

10.4 Room acoustic responses using the image-source method

Using the image-source method (ISM) [155] (see sect. §4.3), a $3 \times 4 \times 2.5$m room (width $\times$ depth $\times$ height) is simulated at 2kHz. A microphone is placed at a distance of 1.2m from the West wall and 1m away from the South wall at an elevation of 1.3m. The sound source is placed at a distance of 2m from the West wall and 3m from the South wall at an elevation of 1.7m. A reverberation time of $T_{60} = 0.3$ is assumed. The simulated response is excited by white Gaussian noise (WGN) of $N = 10,000$ samples. For an increasing channel order $p = 300, 301, \ldots, 2000$, the AR parameters and their corresponding mean squared error (MSE) are calculated using the covariance method. Using the MSE, Akaike's information criterion (AIC) is computed as discussed in

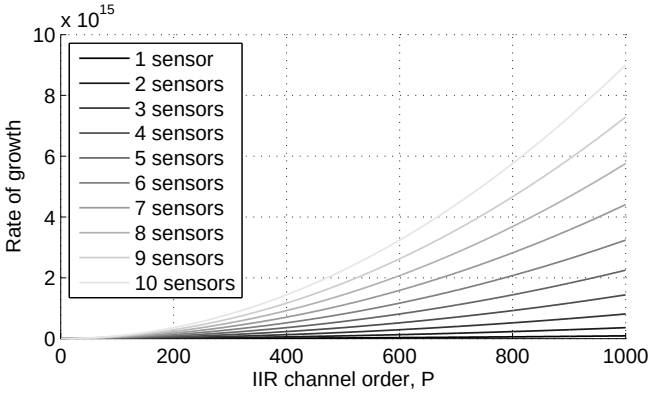


Figure 10.5: *Exponential increase of the computational complexity with the channel order,* P*, in eqn.* (10.19) *with the number of particles for a single sensor (—) up to* 10 *microphones (—) using a source order of* $Q = 15$ *and* $T = 30,000$ *samples.*

sect. §4.4.2 on page 73. By computing the minimum AIC over all model orders, p, the optimal channel order for modelling the simulated ISM response is identified. The experiment is performed for increasing sampling frequencies between $f_s = 2$kHz and $f_s = 16$kHz in 2kHz increments. The optimal channel order in the AIC sense with increasing sampling frequencies is displayed in Fig. 10.4. The optimal channel order for the RIR thus ranges between approximately 350 and 850 parameters depending on the sampling frequency the room acoustics are simulated at.

Assuming $Q = 15$ source parameters and $30,000$ samples, the experiment over an increasing number of microphones in sect. §10.2 is repeated for an increasing channel order, i.e., eqn. (10.19) is evaluated for $P = 1, \ldots, 1000$ parameters and one to ten sensors. The resulting growth rates are plotted in Fig. 10.5. Results indicate that the complexity for $P > 400$ as required for realistic RIRs at $f_s > 4$kHz increases to prohibitive values for real-time implementation of the RBPF.

10.5 Discussion

This chapter investigated the computational complexity of the RBPF. The rate of growth of the algorithm was found to increase quadratically in the channel order and quadratically / cubically in the number of sensors. Although it is, in principle, possible to perform blind dereverberation using channel orders of up to $P = 850$ channel parameters, significant computational burden is therefore induced due to the necessity to evaluate and store the $MP \times 1$ channel parameters and their $MP \times MP$ covariance matrix. The accuracy in modelling the RIR is therefore traded off against the computational burden induced. Furthermore, as dereverberation can be improved by using multiple sensors, a tradeoff between the dereverberation performance and computa-

tional complexity has to be made.

Although the complexity of the RBPF might be prohibitive for a computationally efficient *fullband* implementations, the computational burden can be relaxed by reducing the channel order. Multirate (or subband) filtering approaches can be exploited in order to reduce the channel order, thereby improving computational and memory expense and allowing for more realistic processing of the signal as further discussed in sect. §11.3.

Multirate processing of speech has become a popular approach in the literature due to the severe computational savings that can be obtained. Blind speech dereverberation algorithms using sub-band filtering can be found in, e.g., the work by Enemal and Moonen [241], Gaubitch [36], Daly and Reilly [242], or Hopgood [166, 223].

Chapter 11

Conclusions and future work

This final chapter discusses the closing arguments for this dissertation. Sect. §11.1 summarises the objectives and problems addressed in this thesis. Sect. §11.2 highlights the core results and contributions of this dissertation. An outlook for future research is provided in sect. §11.3.

11.1 Summary and contributions

This dissertation was concerned with the problem of blind dereverberation of speech from stationary and moving speakers using a single and multiple sensors. It was highlighted in an extensive literature review in **Chap. 2** that existing approaches provide valuable insight into the field but also suffer from various individual shortcomings. Nonetheless, this dissertation outlined that most approaches in the literature *in general* are *dictated* by their underlying models and therefore suffer from rigorous assumptions that constrain the methods to very specific subproblems of blind speech dereverberation. The aim of this dissertation was therefore the development of a *general, flexible, and extendible framework* for blind speech dereverberation, allowing for the incorporation of various models for the speech production mechanism as well as different models for RIRs.

Bayesian methods (**Chap. 5**) were investigated for the development of the dereverberation framework. Bayesian methods *benefit* from prior information that is available *a prior* about the speech production mechanism and reverberant distortion. Therefore, a general system model was devised in **Chaps. 3 and 4**. In this model, the speech production mechanism was formulated assuming a TVAR speech model describing the vocal tract in terms of a concatenation of acoustic tubes. The reverberant channel was modelled by an all-pole filter approximating the room transfer function (RTF) according to the solution of the wave equation. For both the speech and channel model,

parameter models are required to specify the exact dynamic of the resulting signals. For the derivation of the proposed framework, these parameter models remained unspecified, such that any suitable model can be incorporated in the algorithm.

Based on the general system model, a novel blind speech dereverberation algorithm using a RBPF was derived in **Chap. 6**. In this framework, the source signal and channel are directly estimated using their optimal estimator, the Kalman filter, and integrated in a sequential importance resampling (SIR) particle filter framework for estimation of the remaining model parameters based on the idea of Rao-Blackwellisation of estimators. The resulting framework therefore facilitates sequential processing, direct source signal estimation, and *blind* channel estimation. Assuming the availability of sufficient processing power, sequential processing could facilitate real-time blind speech dereverberation in the future. As the source signal is estimated directly, channel inversion for the construction of an equalising filter and its associated problems of, e.g., scaling of errors are avoided. Furthermore, as the channel is estimated blindly, prior information about the RIR, such as the T_{60} time, is not required. Moreover, as the channel is estimated using its optimal estimator, issues of implementation within a particle filter due to the enforcement of dynamics on static channels for stationary speakers are avoided. As the observation model is phrased in general terms, the RBPF can be used for single-sensor and multi-sensor blind dereverberation. Whilst single-sensor processing becomes an important aspect for applications where sensor arrays are unfeasible due to their physical size. Multiple sensor can be used for improved dereverberation performance as spatial diversity can be exploited where the physical size of sensor arrays is not of concern.

Observing that the TVAR parameters of speech vary rapidly and relatively smoothly with time, a dynamic TVAR source parameter model was incorporated in the RBPF in **Chap. 7**. In this model, the source parameters are assumed to vary according to a random walk constrained to the area within the unit circle. Experimental results demonstrated that improved speech quality for *unvoiced phonemes*, i.e., stop consonants and fricatives.

In order to improve upon the performance of voiced phonemes, i.e., vowels and semivowels, the source model was extended to a novel PFS based model in **Chap. 8**. In this model, the partial correlation (PARCOR) coefficients of several resonator circuits connected in parallel are assumed to vary according to a random walk. In order to enforce resonant frequencies and bandwidths in the model, the PARCOR samples were constrained to the area corresponding stable parameters and valid frequencies. The corresponding investigations gave valuable insight into the relationships between the

resonant frequencies and bandwidths with the TVAR parameters, their corresponding poles and the PARCOR coefficients. Results on real speech confirmed improved dereverberation for *vowels* as well as stop consonants and fricatives. The PFS based model therefore allows for the incorporation of a speech system model in the RBPF. Furthermore, due to their relation to the reflections of propagating waves in the vocal tract, parameterisation in terms of the PARCOR coefficients adds an additional physical perspective to the models.

By appropriate implementation of source models, the RBPF can therefore model *different classes of phonemes* and therefore relaxes restrictions to specific types of speech sounds often encountered in the literature. Furthermore, **Chap. 9** demonstrated that the stationary speaker scenario can be easily extended to *moving speakers* by appropriate channel modelling. The variation of RIRs with changing source-sensor positions was investigated based on simulated and measured RIRs. Based on the findings, the time-varying channel was modelled as a linear combination of known, time-varying basis functions with unknown, time-invariant channel parameters. Experimental results demonstrated accurate estimation of the anechoic speech signal in segments of sufficient signal power.

Last but not least, Chap. 10 evaluated the *computational complexity* of the RBPF. Although the proposed approach increases linearly in rate of growth with the source model order, the channel order and number of sensors cause quadratic growth in computational complexity.

11.2 Contributions

To summarise, this dissertation made the following contributions:

- Development of a the ***flexible, extendible, and general*** framework for blind speech dereverberation using Bayesian methods, facilitating:
- ***Direct, optimal source signal estimation***, avoiding channel inverse filtering or speech synthesis;
- ***Direct, optimal channel estimation***, avoiding importance sampling of stationary channel parameters and requirements of *a priori* information about the RIR;
- The potential for ***real time processing*** assuming resolution of the computational burden;
- ***Single- and multi-sensor dereverberation***, accommodating physically constrained scenarios where only a single sensor can be utilise, but allowing for exploitation of spatial diversity and hence improvement of dereverberation where physical size of sensor arrays of no concern;

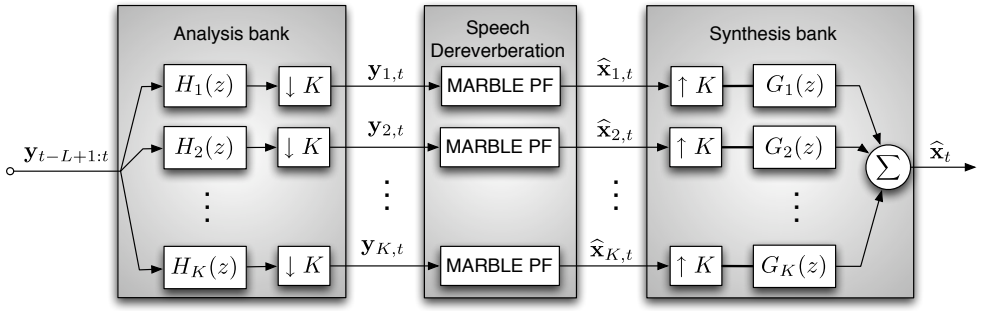


Figure 11.1: *Sequential multirate filtering of fullband reverberant speech,* $\mathbf{y}_{1:t}$*, into* K *subband signals,* $\mathbf{y}_{k,t}$*,* $k \in \mathcal{K}$ *using multirate analysis and synthesis banks with filter length* L *of the dereverberated speech estimated,* $\hat{\mathbf{x}}_t$

- ***Extendibility to a variety of source models*** allowing for voiced, unvoiced, and transient sound modelling; as well as
- Dereverberation of speech from ***stationary and moving speakers***.

Shortcomings and resolutions for future research are discussed in the following, final section of this dissertation.

Overall, a solid and mathematically sound framework of blind signal estimation from noise and distorting channel filters was therefore developed. Future extensions will facilitate more efficient estimation and applicability to a multitude of problems.

11.3 Open extensions and future work

11.3.1 Reducing the computational complexity using multirate filterbanks

A main concern of the proposed approach is the computational complexity discussed in Chap. 10. As mentioned in sect. §10.5, the RBPF could be embedded in a *multirate* filtering approach to reduce the computational burden. In this framework, an analysis filter bank consisting of K filters $H_k(z)$ $k \in \mathcal{K}$, channelizes the input signal, $\mathbf{y}_{1:t}$, into K sub-band signals, $\mathbf{y}_{k,t}$, and decimates the resulting signals by a factor of K (denoted as $\downarrow K$). Because of the reduced sampling frequency, less model parameters and samples are required, leading to more efficient and faster processing. After processing, the estimated subband signals, $\mathbf{x}_{k,t}$ are recombined by interpolating by a factor of K (denoted as $\uparrow K$) and applying K synthesis filters, $G_k(z)$. The fullband signal is the sum over the output of the synthesis filters. A sequential multirate processing filterbank is shown in Fig. 11.1.

Classic perfect reconstruction (PR) filterbanks assume that each subband signal is extracted using what can essentially be considered as a bandpass filter. In practice, the bandpass analysis filters have non-zero stop-band gain. Thus, the sub-band signals

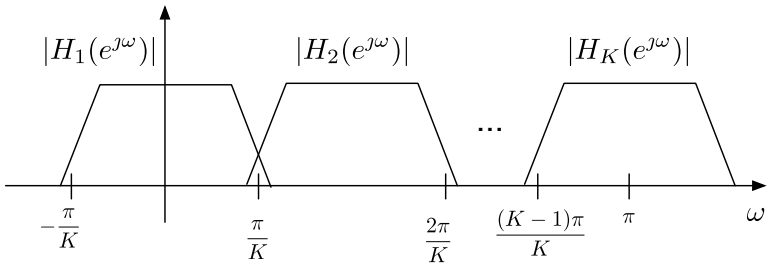


Figure 11.2: K*-channel analysis filter*

are not strictly band-limited and the responses of subband signals overlap. Each signal can have energy for bandwidths exceeding the ideal passband region as illustrated in Fig. 11.2. Hence, aliasing is introduced by decimation [231]. PR filterbanks permit aliasing in the analysis bank and cancel the alias component using appropriate synthesis filters. This approach is appropriate in systems where any subband processing is absent, i.e., where the analysis and synthesis filter are connected back to back and the input signal is equal to the output signal. PR subsampling is thus undesirable for systems that contain subband processing blocks sensitive to aliasing, such as adaptive filtering [237,243], or, in this case, blind speech dereverberation. Instead, near perfect reconstruction oversampled filterbanks *suppress* aliasing in the subbands rather than *cancelling* aliasing at the output. generalised discrete Fourier transform (GDFT) filterbanks [234,244–246] are particularly promising candidates suitable for subband signal processing applications as discussed extensively in, e.g., [233–236].

11.3.2 Hybrid speech model using a Markov switching model

It is desirable to *combine* the dynamic TVAR parameter model for unvoiced speech with the PFS model for voiced and transit sounds, and possibly even extend the model to accommodate for a sinusoidal model for improved modelling of vowels and semivowels. Rather than choosing a single model that best represents all modes of speech, different speech models can be combined using a model-switching regime. Multiple model particle filters facilitate source signal estimation using J models that can transit from one to another. multiple model (MM) particle filters consist of a bank of J filters for each model for source signal estimation combined with a filter for parameter estimation [17]. The source signal is generalised to the form

$$\mathbf{x}_t = f\left(\mathbf{x}_{t-1},\, \mathbf{a}_t^j,\, \mathcal{M}_t\right) + \mathbf{D}_t\,\mathbf{v}_t \qquad\qquad v_t \sim \mathcal{N}\left(\mathbf{0}_{Q_t\times 1},\, \mathbf{I}_{Q_t}\right) \tag{11.1}$$

where $f(\cdot)$ is a function of the previous source signal samples, $\mathbf{x}_{t-1}$, the source model parameters, $\mathbf{a}_t$, of the current model, $\mathcal{M}_t$, of model order Q_t. The MM particle filter is governed by a mode regime variable, m_t, determining the which of the three dy-

namic models is currently in effect. The mode transition is modelled by a Markovian switching model, e.g., using a Markov chain, i.e., [247]

$$\Pr\left(M_t = j \,\middle|\, M_{t-1} = i\right) = \pi_j \qquad\qquad j \in \mathcal{J}. \tag{11.2}$$

where $\{\pi_{kj}\}_{j\in\mathcal{J}}$ denotes the transitional switching probabilities for the J model candidates. Recalling that in Chap. 6 the dynamic model parameters were given by $\theta_{0:t}$, the dynamic parameter space is augmented to include the mode regime variable, m_t via $\lambda \triangleq \begin{bmatrix}\theta_{0:t}^T & m_t\end{bmatrix}$. Assuming that N particles of the parameter space are drawn each particle, $\lambda^{(i)}$, each particle's weight is (still) denoted as $w_t^{(i)}$, where $i \in \mathcal{N}$.

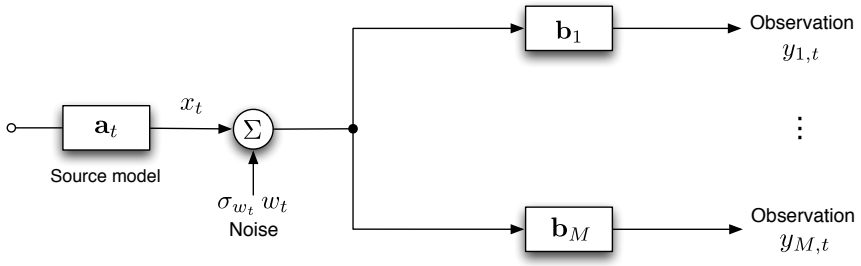
Estimation of the model regime variable itself could be improved by utilising a voice activity detector (VAD) in order to distinguish between different types of phonemes. Lehmann and Johansson [248] implement a VAD based particle filter for target tracking in order to avoid misguidance of the tracker due to silence gaps. The VAD is based on the signal-to-noise ratios (SNRs) of data: In segments of voice activity, the signal power is assumed to be significantly larger than the noise power. Thus, if the SNR lies below a certain threshold, the signal is flagged as silent. The voice-activity pdf is thus integrated in the particle filter, such that the estimator can switch between appropriate tracking for voice active and inactive speech segments.

Although a voice activity detector based on an SNR threshold does not appear appropriate for the distinction of phonemes, a VAD based particle certainly would proof feasible for a multiple model switching particle filter. Stop consonants and fricatives are often semivowels and thus partially unvoiced. Thus, introducing a dependency of the Markovian switching model on a VAD distinguishing between voiced, partially voiced, and unvoiced phonemes would allow for switching between the TVAR model for unvoiced speech, the PFS-PARCOR model for transient sounds, and a harmonic model for vowels.

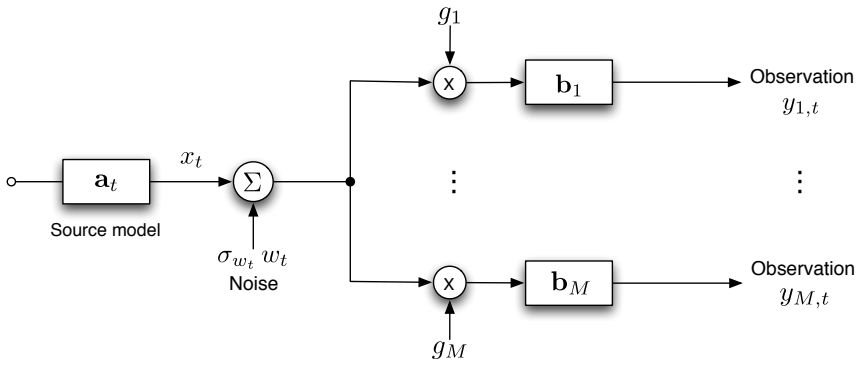
Implementation of a VAD would also allow for the exclusion of signal segments with low signal power, avoiding issues for blind dereverberation for moving speakers in Chap. 9.

11.3.3 Inclusion of channel gain terms for multiple speakers

The system model proposed in eqn. (6.9) on page 115 and summarised in Fig. 11.3a facilitates multiple observations as demonstrated by the experiments in Chap. 7. Generally, a channel gain term should added at the system output if the parameters of the



(a) *Ignoring channel gain terms.*



(b) *Considering individual channel gain terms.*

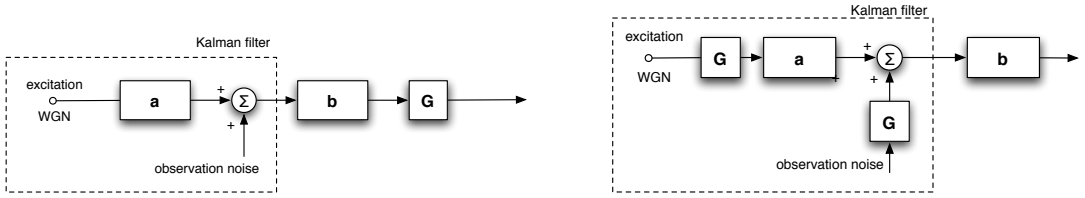
Figure 11.3: *Comparison of MIMO setup ignoring and considering individual channel gain terms.*

system are known (Fig. 11.4a). The system model can be extended to account for individual channel gains by introducing the gain factor matrix, $\mathbf{G} \triangleq \text{diag}\left[\mathrm{g}_1 \quad \dots \quad \mathrm{g}_M\right]$ where $\{\mathrm{g}_m\}_{m\in\mathcal{M}}$ is the gain factor for the channel between the source and the m^{th} microphone, such that the observation space becomes

$$\mathbf{y}_t = \mathbf{Y}_{t-1}\mathbf{b} + \mathbf{G}\left[\mathbf{C}^{\mathsf{T}}\mathbf{x}_t + \mathbf{\Sigma}_{\mathbf{w}_t}\,\mathbf{w}_t\right] \tag{11.3}$$

as illustrated in Fig. 11.3b. For a single sensor, this setup corresponds to the system in Fig. 11.4a and is equivalent to accounting for the channel gain at the source input and at the input of the observation noise (Fig. 11.4b). If the system parameters are unknown, the gain term is implicitly included in the estimation of the parameters. Hence, for single sensors, the channel gain can be omitted due to a scaling ambiguity.

For multiple sensors, the scaling gain cannot be rewritten similar to the principle in Fig. 11.4b. The channel gain term, $\mathbf{G}$, should therefore be explicitly included in the system model. Assuming that $\mathbf{G}$ is static, i.e., does not vary with time, it cannot be estimated using importance sampling. Similar to the channel parameters in Chap. 6, its optimal estimator therefore need to be derived. However, as the gain is involved in the observation space in eqn. (11.3), the estimator of the source signal and channel



(a) *General reverberant system with channel gain at output*

(b) *Equivalent system to Fig. 11.4a with channel gain at source and noise input*

Figure 11.4: *Reverberant system including gain term illustrating equivalent arrangements leading to scaling ambiguity*

parameters are dependent on the channel gain and hence need to be rederived as well. This is easily shown by applying the Rao-Blackwellisation principle in sect. §5.6 again: Assuming that $\mathbf{z}_{0:t} \triangleq \begin{bmatrix} \mathbf{x}_{0:t}^{\mathsf{T}} & \mathbf{b}^{\mathsf{T}} & \mathbf{g}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$, the minimum mean-square error (MMSE) estimator of all three variables can be found by evaluating

$$\widehat{\mathbf{z}}_{0:t} = \begin{bmatrix} \int_{\mathbb{R}^Q} \mathbf{x}_{0:t}\, p\left(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right) d\mathbf{x}_{0:t} \\ \int_{\mathbb{R}^P} \mathbf{b}\, p\left(\mathbf{b} \mid \mathbf{y}_{1:t}\right) d\mathbf{b} \\ \int_{\mathcal{G}} \mathbf{g}\, p\left(\mathbf{g} \mid \mathbf{y}_{1:t}\right) d\mathbf{g} \end{bmatrix} = \begin{bmatrix} \widehat{\mathbf{x}}_{0:t} \\ \widehat{\mathbf{b}} \\ \widehat{\mathbf{g}} \end{bmatrix} \tag{11.4}$$

where $p\left(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right)$ is the marginalised posterior pdf of the source signal, independent of the channel and gain factor, and $p\left(\mathbf{b} \mid \mathbf{y}_{1:t}\right)$ is the marginalised channel posterior pdf, independent of the channel gain, i.e.,

$$p\left(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}\right) = \int \left[\int p\left(\mathbf{x}_{0:t} \mid \mathbf{y}_{1:t}, \mathbf{b}, \mathbf{g}\right) p\left(\mathbf{b} \mid \mathbf{y}_{1:t}, \mathbf{g}\right) d\mathbf{b} \right] p\left(\mathbf{g} \mid \mathbf{y}_{1:t}\right) d\mathbf{g} \tag{11.5}$$

$$p\left(\mathbf{b} \mid \mathbf{y}_{1:t}\right) = \int p\left(\mathbf{b} \mid \mathbf{y}_{1:t}, \mathbf{g}\right) p\left(\mathbf{g} \mid \mathbf{y}_{1:t}\right) d\mathbf{g}. \tag{11.6}$$

Not only do the channel parameters need to be marginalised from the source signal posterior pdf, also does the channel gain need to be marginalised from both the source signal and channel posterior pdf. Unfortunately, this derivation would require a significant amount of work beyond the timeframe of this thesis.

11.3.4 Model order selection using a JMS

Undermodelling of the channel leads to significant performance deterioration due to omitting high-energy taps in the channel filter. As the channel order is unknown in practice and trial-and-error simulations until a suitable channel is determined are gen-

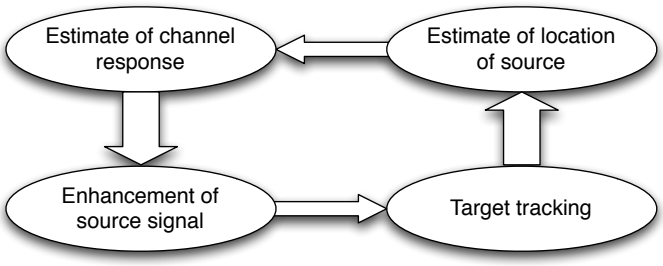


Figure 11.5: *Joint dependency between enhancement and tracking*

erally unfeasible, the implementation of model selection order schemes are desirable.

In scenarios where the speaker position is changing, the channel model is consistently varying with time, which can lead to the birth and death of channel poles. Hence, the channel order is varying with time as well. jump Markov system (JMS) are particularly well suited for model order selection schemes where the model order is changing with time. Defining the state of the JMS as model order, P_t, at time t, the observation at microphone $m \in \mathcal{M}$ can be written as

$$y_{m,t} = \sum_{p \in \mathcal{P}_t} b_{m,p}\, y_{m,t-p} + x_t + \sigma_{m,w_t}\, w_{m,t}. \tag{11.7}$$

Note that eqn. (11.7) differs from the observation model in eqn. (4.15) on page 77 by the limit of the sum, now ranging between 1 and time-varying P_t rather than constant P. The model order changes (or jumps) with transition probability, $p(P_t \mid P_{t-1})$, dependent on its previous state, P_{t-1}. As P_t is time-varying and can be assigned to obey a first-order Markov chain, it makes sense to track the model order in the importance sampling step of the particle filter. In this case, the unknown channel order is appended to the space of time-varying model parameters, $\boldsymbol{\theta}_t = \begin{bmatrix} \mathbf{a}_t^T & \boldsymbol{\phi}_{\mathbf{w}_t} & \phi_{\nu_t} & P_t \end{bmatrix}^T$. This principle was successfully implemented for tracking harmonic components in music and speech in [249], where the number of components changes in time.

11.3.5 Joint tracking and enhancement

In order to improve upon source signal estimation, particularly for moving speakers, the model of the surrounding room acoustics should be improved. In order to obtain a better channel model, positional knowledge, and hence tracking, of the sound source would be advantageous. Should the object be localised or tracked, a clean signal is necessary in order to track the true source rather than the reflections of source signal. Target tracking and signal enhancement are thus jointly dependent as illustrated in Fig. 11.5.

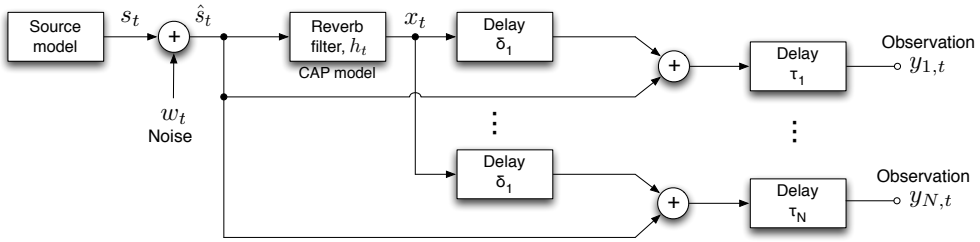


Figure 11.6: *System model facilitating RIR to be implicitly dependent on the source position through a function of delays.*

Target tracking – an example of remote source sensing – and audio enhancement are both widely researched topics with numerous applications. Various methods for the solution of either individual problem exist. However, the combination of simultaneous tracking and enhancement has received little attention in the literature so far. Therefore, the establishment of a joint framework is of practical and theoretical interest to both the research community and industry. For example, this problem finds application in the civilian and military sectors such as surveillance, command and control, air traffic control and navigation.

The models proposed in this thesis can be extended to facilitate joint source tracking and dereverberation by modelling the RIR as a function of the location of the source position, relative to the sensor as shown in Fig. 11.3.5.

Part IV

Appendices

Appendix A

Background derivations

A.1 Weighted RLS approach

This section shows how the algorithm in [57] can be rephrased as a Kalman filter as mentioned in sect. §2.3 on page 18. Yoshioka *et al.* propose a sequential algorithm in [57] that operates in the short-time Fourier transform (STFT) domain and is based on Bayesian updates of the reverberant algorithm. The STFT, $\mathbb{Y}_{m,t,\omega}$, of the observed signal, $y_{m,t}$, at time t observed at microphone $m \in \mathcal{M}$ in frequency bin $\omega \in \Omega$ is constructed from its time-domain version via

$$\mathbb{Y}_{m,\tau,\omega} = \text{STFT}\left[y_{m,t}\right] = \sum_{t=-\infty}^{\infty} y_{m,t}\, w_{t-\tau}\, e^{-\jmath\omega t} \tag{A.1}$$

where w_t is a window, typically chosen as the Hann window or the Hamming window, which is defined as

$$w_t = 0.54 - 0.46\cos\left\{\frac{2\pi t}{N-1}\right\} \tag{A.2}$$

where N is the window length. Based on AR modelling of the observed signal using multiple microphones, the output at the first microphone in the STFT domain can be expressed in terms of the reverberant filter as

$$\mathbb{Y}_{1,\tau,\omega} = \sum_{m\in\mathcal{M}} \sum_{p\in\mathcal{P}_\omega} b^{\star}_{m,p,\omega}\, \mathbb{Y}_{m,\tau-p,\omega} + \mathbb{X}_{\tau,\omega} = \mathbf{b}_\omega^{H}\, \underline{\mathbb{Y}}_{\tau-1,\omega} + \mathbb{X}_{\tau,\omega} \tag{A.3}$$

where $\star$ denotes the complex conjugate and H denotes the Hermitian, $X_{\tau,\omega}$ is the STFT of the source signal in frame τ and frequency band $\omega \in \Omega$; P_ω is the AR model and $\mathbf{b}_\omega = \begin{bmatrix} b_{1,1,\omega} & \dots & b_{1,P,\omega} & \dots & b_{M_\omega,1,\omega} & \dots & b_{M_\omega,P,\omega} \end{bmatrix}^{T}$ are the coefficients of the infinite impulse response (IIR) channel model in the ω^{th} frequency band;

and $\mathbf{Y}_{\tau-1,\omega} = \begin{bmatrix} Y_{1,\tau-1,\omega} & \dots & Y_{1,\tau-P_\omega,\omega} & \dots & Y_{M,\tau-1,\omega} & \dots & Y_{M,\tau-P_\omega,\omega} \end{bmatrix}^T$ are the past observed samples. Eqn. (A.3) can be rearranged for the source signal as

$$\mathbb{X}_{\tau,\omega} = Y_{1,\tau,\omega} - \mathbf{b}_\omega^H \underline{\mathbb{Y}}_{\tau-1,\omega}. \tag{A.4}$$

Thus, the task in [57] is to identify the channel coefficients, $\mathbf{B} \triangleq \{\mathbf{b}_\omega\}_{\omega\in\Omega}$ given the observations $\overline{\underline{\mathbb{Y}}}_{0:t} = \left\{\underline{\mathbb{Y}}_{\tau-1,\omega}\right\}_{\omega\in\Omega}$. Using Bayes's rule, the posterior pdf of the channel coefficients is expressed as

$$p\left(\mathbf{B} \mid \overline{\underline{\mathbb{Y}}}_{0:\tau}\right) = \frac{p\left(\underline{\overline{\mathbb{Y}}}_\tau \,\middle|\, \overline{\underline{\mathbb{Y}}}_{0:\tau-1}, \mathbf{B}\right) p\left(\mathbf{B} \mid \overline{\underline{\mathbb{Y}}}_{0:\tau-1}\right)}{\int p\left(\underline{\overline{\mathbb{Y}}}_\tau \,\middle|\, \overline{\underline{\mathbb{Y}}}_{0:\tau-1}, \mathbf{B}\right) p\left(\mathbf{B} \mid \overline{\underline{\mathbb{Y}}}_{0:\tau-1}\right) d\mathbf{B}} \tag{A.5}$$

If the source signal, $\mathbb{X}_{\tau,\omega}$, is assumed to be Gaussian with with zero mean and covariance $\sigma^2_{\nu_{\tau,\omega}}$, where $\sigma^2_{\nu_{\tau,\omega}}$ corresponds to the short-time power spectrum of the speech signal, then the likelihood of the observations, $p\left(\underline{\overline{\mathbb{Y}}}_\tau \,\middle|\, \overline{\underline{\mathbb{Y}}}_{0:\tau-1}, \mathbf{B}\right)$ can be expressed as

$$p\left(\underline{\overline{\mathbb{Y}}}_\tau \,\middle|\, \overline{\underline{\mathbb{Y}}}_{0:\tau-1}, \mathbf{B}\right) = \prod_{\omega\in\Omega} \mathcal{N}\left(\mathbb{Y}_{1,\tau,\omega} \,\middle|\, \mathbf{b}_\omega^H \underline{\mathbb{Y}}_{\tau-1,\omega},\, \sigma_{\nu_{\tau,\omega}}\right). \tag{A.6}$$

If it can be assumed that the posterior pdf of the channel after observing $\overline{\underline{\mathbb{Y}}}_{0:\tau-1}$ is Gaussian with mean $\boldsymbol{\mu}_{\tau-1,\omega}$ and covariance $\boldsymbol{\Sigma}_{\tau-1,\omega}$, then the posterior at τ is is expressed as

$$p\left(\mathbf{B} \mid \overline{\underline{\mathbb{Y}}}_{0:\tau}\right) = \prod_{\omega\in\Omega} \mathcal{N}\left(\mathbf{b}_\omega \,\middle|\, \boldsymbol{\mu}_{\tau,\omega},\, \boldsymbol{\Sigma}_{\tau,\omega}\right) \tag{A.7}$$

where the mean and covariance are given as

$$\boldsymbol{\mu}_{\tau,\omega} = \boldsymbol{\Sigma}_{\tau,\omega}\left(\frac{\underline{\mathbb{Y}}_{\tau-1,\omega}\, \mathbb{Y}^\star_{\tau,\omega}}{\sigma^2_{\nu_{\tau,\omega}}} + \boldsymbol{\Sigma}^{-1}_{\tau-1,\omega}\, \boldsymbol{\mu}_{\tau-1,\omega}\right) \tag{A.8}$$

$$\boldsymbol{\Sigma}_{\tau,\omega} = \left(\frac{\underline{\mathbb{Y}}_{\tau-1,\omega}\, \underline{\mathbb{Y}}^H_{\tau-1,\omega}}{\sigma^2_{\nu_{\tau,\omega}}} + \boldsymbol{\Sigma}_{\tau-1,\omega}\right)^{-1} \tag{A.9}$$

In order to avoid computation of the inverse matrix of order P_ω, the algorithm is simplified by application of the Woodbury matrix identity to give

$$\boldsymbol{\mu}_{\tau,\omega} = \boldsymbol{\mu}_{\tau-1,\omega} + \mathbf{k}_{\tau,\omega}\, \hat{\mathbb{X}}^\star_{\tau,\omega} \tag{A.10a}$$

$$\boldsymbol{\Sigma}_{\tau,\omega} = \frac{1}{\alpha}\left(\boldsymbol{\Sigma}_{\tau-1,\omega} - \mathbf{k}_{\tau,\omega}\, \underline{\mathbb{Y}}^H_{\tau-1,\omega}\, \boldsymbol{\Sigma}_{\tau-1,\omega}\right) \tag{A.10b}$$

where $0 \leq \alpha \leq 1$ is a forgetting factor and $\mathbf{k}_{\tau,\omega}$ is the gain factor:

$$\mathbf{k}_{\tau,\omega} = \frac{\boldsymbol{\Sigma}_{\tau-1,\omega}\underline{\mathbb{Y}}_{\tau-1,\omega}}{\alpha\,\sigma^2_{v_{\tau,\omega}} + \underline{\mathbb{Y}}^{H}_{\tau-1,\omega}\,\boldsymbol{\Sigma}_{\tau-1,\omega}\,\underline{\mathbb{Y}}_{\tau-1,\omega}} \tag{A.10c}$$

A.2 Derivation of Webster's equation

This section outlines the derivation of Webster's equation as first mentioned sect. §3.2 on page 32. For completeness, recall that Newton's laws of force describe the dependency of the speed of sound on the density variation with pressure, i.e.,

$$\frac{1}{\rho c^2}\frac{\partial p(x,t)}{\partial t} + \operatorname{div}(\mathbf{v}) = 0 \tag{A.11a}$$

$$\rho\frac{\partial \mathbf{v}}{\partial t} + \nabla(p(x,t)) = 0, \tag{A.11b}$$

where $p(x,t)$ is the sound pressure dependent on the distance, x, and time, t, $\mathbf{v}$ is the vector velocity of an air particle, ρ is the density of air in the tube, and c is the speed of sound. For one-dimensional, planar airflow, the volume velocity is generally used instead of the particle velocity, where $u = Av$, with $u(x,t)$ as the volume velocity, and $A(x)$ is the vocal tract area as a function of the distance is the sound pressure. Thus, eqn. (A.11) reduces to the continuity of mass and momentum equations which are given respectively by

$$\begin{aligned} -\frac{\partial u(x,t)}{\partial x} &= \frac{1}{\rho c^2}\frac{\partial (p(x,t)\,A(x))}{\partial t} + \frac{\partial A(x)}{\partial t} \\ -\frac{\partial p(x,t)}{\partial x} &= \rho\frac{\partial (u(x,t)/A(x))}{\partial t}. \end{aligned} \tag{A.12}$$

as $A(x)$ is independent of t,

$$\begin{aligned} -\frac{\partial u(x,t)}{\partial x \partial t} &= \frac{A(x)}{\rho c^2}\frac{\partial^2 p(x,t)}{\partial t^2} \\ -\frac{\partial u(x,t)}{\partial t \partial x} &= \frac{1}{\rho}\frac{\partial}{\partial x}\left[A(x)\frac{\partial p(x,t)}{\partial x}\right] \end{aligned} \tag{A.13}$$

which can be combined by equating the right hand side (RHS) to obtain Webster's equation as in eqn. (3.1).

A.3 Lossless acoustic tube model

A.3.1 Reflection of sound waves

This section derives the reflection coefficients of the acoustic tube model in sect. §3.2.1 on page 34. Assuming K uniform sections of the acoustic tube, where the cross-sectional area $A_k(x)$ of the k^{th} tube, $k \in \mathcal{K}$, is constant over the total length of each section, Webster's horn equation reduces to

$$\frac{1}{c^2}\frac{\partial^2 p_k(x,t)}{\partial t^2} = \frac{\partial^2 p_k(x,t)}{\partial x^2}. \tag{A.14}$$

where $p_k(x,t)$ is the sound pressure in the k^{th} tube. Similarly, eqn. (A.14) can be constructed in terms of the volume velocity, $u(x,t)$, similar to the pressure, $p(x,t)$, which, assuming constant area in each section, reduces to,

$$\frac{1}{c^2}\frac{\partial^2 u_k(x,t)}{\partial t^2} = \frac{\partial^2 u_k(x,t)}{\partial x^2}. \tag{A.15}$$

where $u_k(x,t)$ is the volume velocity in the k^{th} tube. Eqn. (A.14) and (A.15) can be expressed as a linear combination of forward and reverse travelling waves as illustrated in Fig. 3.2. The forward waves move from the glottis towards the lips, whilst the reverse waves travel from the lips in the direction of the glottis. Thus, eqns. (A.14) and (A.15) become [109, ch. 4.2.1, p. 63ff.]

$$p_k(x,t) = p_k^+(x, t - x/c) - p_k^-(x, t + x/c) \tag{A.16a}$$
$$u_k(x,t) = u_k^+(x, t - x/c) - u_k^-(x, t + x/c). \tag{A.16b}$$

Inserting eqn. (A.16) into eqn. (A.11), and applying the relationship [109]

$$\frac{\partial f(t \pm x/c)}{\partial t} = \pm c\frac{\partial f(t \pm x/c)}{\partial x}, \tag{A.17}$$

for any function $f(\cdot)$, the pressure can be expressed in terms of the velocity as

$$p_k(x,t) = \frac{\rho c}{A_k}\left[u_k^+(x, t - x/c) + u_k^-(x, t - x/c)\right]. \tag{A.18}$$

Due to conservation of volume continuity in both time and space, at the boundary between section k and $k+1$ the following condition needs be fulfilled:

$$u_k^+(\ell_k, t - \tau_k) - u_k^-(\ell_k, t + \tau_k) = u_{k+1}^+(0,t) - u_{k+1}^-(0,t), \tag{A.19}$$

where ℓ_k, $k \in \mathcal{K}$ is the length of the k^{th} section and $\tau_k \triangleq \ell_k/c$ for clarity. Likewise, eqn. (A.18) can be rewritten as

$$\frac{\rho c}{A_k}\left[u_k^+(\ell_k, t-\tau_k) + u_k^-(\ell_k, t+\tau_k)\right] = \frac{\rho c}{A_{k+1}}\left[u_{k+1}^+(0,t) + u_{k+1}^-(0,t)\right] \tag{A.20}$$

Thus, when a wave front meets the discontinuity area of a section, part of the wave propagates through to the next section, whilst the remainder is reflected back into its own section. A wave will only propagate fully if the impedance of the next section meets that of the previous section, i.e., the cross-sectional areas $A_k = A_{k+1}$ [109]. This concept is clarified by solving for $u_{k+1}^+(0,t)$ and $u_k^-(\ell_k, t+\tau_k)$, i.e.,

$$\begin{aligned} u_{k+1}^+(0,t) &= (1+r_k)\, u_k^+(\ell_k, t-\tau_k) + r_k u_{k+1}^-(0,t) \\ u_k^-(\ell_k, t+\ell_k/c) &= -r_k u_k^+(\ell_k, t-\tau_k) + (1-r_k)\, u_{k+1}^-(0,t), \end{aligned} \tag{A.21}$$

where the reflection coefficient, r_k, can be written in the form of eqn. (3.2).

A.3.2 Transfer function

This section derives the transfer function of the vocal tract in sect. §3.3.1 on page 38. Eqn. (A.21) showed that the volume velocity flow at each junction can be expressed as

$$\begin{aligned} u_{k+1}^+(0,t) &= (1+r_k)\, u_k^+(\ell_k, t-\tau_k) + r_k u_{k+1}^-(0,t) \\ u_k^-(\ell_k, t+\ell_k/c) &= -r_k u_k^+(\ell_k, t-\tau_k) + (1-r_k)\, u_{k+1}^-(0,t), \end{aligned} \tag{A.21}$$

the respective z-transforms are given as

$$U_{k+1}^+(z) = (1+r_k)\, z^{-1/2}\, U_k^+(z) + r_k\, U_{k+1}^-(z) \tag{A.22}$$

$$U_k^-(z) = -r_k\, z^{-1} U_k^+(z) + (1-r_k)\, z^{-1/2} U_{k+1}^-(z) \tag{A.23}$$

or equivalently, by solving for $U_k^+(z)$ and $U_k^-(z)$,

$$U_k^+(z) = \frac{z^{1/2}}{1+r_k} U_{k+1}^+(z) - \frac{r_k\, z^{1/2}}{1+r_k} U_{k+1}^-(z) \tag{A.24}$$

$$U_k^-(z) = \frac{-r_k\, z^{-1/2}}{1+r_k} U_{k+1}^+(z) + \frac{z^{-1/2}}{1+r_k} U_{k+1}^-(z). \tag{A.25}$$

Thus, in matrix form, the flow of the system can be formulated as

$$\mathbf{U}_k = \mathbf{Q}_k\, \mathbf{U}_{k+1} \tag{A.26}$$

where

$$\mathbf{U}_k \triangleq \begin{bmatrix} U_k^+(z) \\ U_k^-(z) \end{bmatrix} \tag{A.27}$$

$$\mathbf{Q}_k \triangleq \begin{bmatrix} \frac{z^{1/2}}{1+r_k} & -\frac{r_k z^{1/2}}{1+r_k} \\ -\frac{r_k z^{-1/2}}{1+r_k} & \frac{z^{-1/2}}{1+r_k} \end{bmatrix} = z^{1/2} \underbrace{\begin{bmatrix} \frac{1}{1+r_k} & -\frac{r_k}{1+r_k} \\ -\frac{r_k z^{-1}}{1+r_k} & \frac{z^{-1}}{1+r_k} \end{bmatrix}}_{\hat{\mathbf{Q}}_k} \tag{A.28}$$

Thus, by iteratively applying eqn. (A.26), the input to the first tube is given as

$$\mathbf{U}_1 = \prod_{k\in\mathcal{K}} \mathbf{Q}_k \mathbf{U}_{K+1} = \prod_{k\in\mathcal{K}} z^{1/2} \hat{\mathbf{Q}}_k \mathbf{U}_{K+1} = z^{K/2} \prod_{k\in\mathcal{K}} \hat{\mathbf{Q}}_k \mathbf{U}_{K+1} \tag{A.29}$$

Also, using eqn. (A.24), the flow at the glottis can be expressed as (EXPLAIN! P. 94 Rabiner)

$$U_G(z) = \frac{2}{1+r_G} U_1^+ - \frac{2r_G}{1+r_G} U_1^- = \begin{bmatrix} \frac{2}{1+r_G} & -\frac{2r_G}{1+r_G} \end{bmatrix} \mathbf{U}_1. \tag{A.30}$$

As the last section, U_{N+1}, is the volume velocity at the lips,

$$\mathbf{U}_{K+1} = \begin{bmatrix} U_L(z) \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} U_L(z), \tag{A.31}$$

such that the inverse transfer function is given as

$$\frac{1}{V(z)} = \frac{U_G(z)}{U_L(z)} = z^{K/2} \begin{bmatrix} \frac{2}{1+r_G} & -\frac{2r_G}{1+r_G} \end{bmatrix} \prod_{k=1}^{K} \hat{\mathbf{Q}}_k \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{A.32}$$

As the elements of $\hat{\mathbf{Q}}_k$ are either constant or proportional to $z^{-1/2}$, the matrix product reduces to a polynomial in z^{-1} of order N. Thus, the transfer for the lossless tube model can be expressed by eqns. (3.5) and (3.6) on page 40.

A.4 PARCOR parameter model

A.4.1 Two-stage lattice structure output

This section derives the relation between the AR parameters and the corresponding PARCOR (or reflection) coefficients of a second-order AR process as first mentioned in sect. §3.4.2 on page 52.

Any IIR filter of order Q can be described by the direct-form structure illustrated in Fig. 3.12a. Alternatively, as shown in [144, ch. 9.3.5], the direct-form IIR filter is

equivalent to a lattice structure illustrated in Fig. 3.12b, where $\psi_{t,q}$, $q \in \mathcal{Q}$ denote the reflection coefficients of the lattice structure. Recalling the relation between the backward and forward lattice structure outputs in eqn. (3.26) and reducing to $Q = 2$ for the case of resonator circuits, the corresponding reflection coefficients, $\lambda_t^{(q)}$ and $\phi_t^{(q)}$ for $q = 1, 2$ simplify to

$$\lambda_{1,t} = \lambda_{0,t} + \psi_{1,t}\,\phi_{0,t-1} = x_t + \psi_{1,t}\,x_{t-1} \tag{A.33a}$$

$$\phi_{1,t} = \psi_{1,t}\,\lambda_{0,t} + \phi_{0,t-1} = \psi_{1,t}\,x_t + x_{t-1} \tag{A.33b}$$

$$\lambda_{2,t} = \lambda_{1,t} + \psi_{2,t}\,\phi_{1,t-1} = x_t + \psi_{1,t}(1 + \psi_{2,t})x_{t-1} + \psi_{t,2}x_{t-2} \tag{A.33c}$$

$$\begin{aligned}\phi_{2,t} &= \psi_{2,t}\,\lambda_{1,t} + \phi_{1,t-1} = \psi_{2,t}\,(x_t + \psi_{1,t}\,x_{t-1}) + \psi_{1,t-1}\,x_{t-1} + x_{t-2} \\ &= \psi_{2,t}\,x_t + (\psi_{1,t}\psi_{2,t} + \psi_{1,t-1})\,x_{t-1} + x_{t-2}\end{aligned} \tag{A.33d}$$

Solving eqn. (A.33c) for x_t,

$$x_t = -\psi_{1,t}\,(1 + \psi_{2,t})\,x_{t-1} - \psi_{t,2}\,x_{t-2} + \lambda_t^{(2)}$$

and hence x_t as described by eqn. (3.27).

A.4.2 Relation between reflection and PARCOR coefficients

This section shows that the reflection coefficients are equivalent to PARCOR coefficients as mentioned in sect. §3.4.2.1 on page 54. Recalling eqn. (3.26b) for the forward stage of the lattice structure:

$$\lambda_{q-1,t} = \lambda_{q,t} - \psi_{q,t}\,\phi_{q-1,t-1} \tag{3.26b}$$

The variance of the forward prediction, $\lambda_{q,t}$, can be expressed as

$$\text{var}\,[\lambda_{q,t}] = \mathbb{E}\left[|\lambda_{q,t}|^2\right] = \mathbb{E}\left[|\lambda_{q-1,t} + \psi_{q,t}\phi_{q-1,t-1}|^2\right] \tag{A.34}$$

$$= \mathbb{E}\left[|\lambda_{q-1,t}|^2\right] + \psi_{q,t}^2\,\mathbb{E}\left[|\phi_{q-1,t-1}|^2\right] + 2\psi_{q,t}\,\mathbb{E}\left[\lambda_{q-1,t}\,\phi_{q-1,t-1}\right]. \tag{A.35}$$

Minimising with respect to the reflection coefficient, $\psi_{q,t}$,

$$\frac{\partial \mathbb{E}\left[|\lambda_{q,t}|^2\right]}{\partial \psi_{q,t}} = 2\psi_{q,t}\,\mathbb{E}\left[|\phi_{q-1,t-1}|^2\right] + 2\,\mathbb{E}\left[\lambda_{q-1,t}\,\phi_{q-1,t-1}\right] = 0$$

such that solving for $\psi_{q,t}$ yields eqn. (3.30).

A.4.3 Transfer function of PARCOR model

This section shows that the transfer function of the PARCOR model can be expressed in terms on a recursion of the denominator as first mentioned in sect. §3.4.2.2 on page 55. Transforming the lattice equations in eqn. (3.26) to the z-domain,

$$\Lambda_0(z) = \Phi_0(z) = V(z) \tag{A.36a}$$

$$\Lambda_q(z) = \Lambda_{q-1}(z) + \psi_{q,t}\, z^{-1}\, \Phi_{q-1}(z) \tag{A.36b}$$

$$\Lambda_Q(z) = N_P(z) \tag{A.36c}$$

where $\Lambda_q(z)$ is the z-transform of $\lambda_{q,t}$ and $\Phi_q(z)$ is the z-transform of $\phi_{q,t}$. As mentioned in sect. §3.4.2.1, the forward and backward feedback path are related via $\phi_{q,t} = \lambda_{q,t-1}$, or in z-domain, $\Phi_q(z) = \Lambda_q(z^{-1})$ [144, ch. 12.3.2]. Inserting for $\Phi_{q-1}(z)$ in eqn. (A.36b) thus yields

$$\Lambda_q(z) = \Lambda_{q-1}(z) + \psi_{q,t}\, z^{-1}\, \Lambda_{q-1}(z^{-1}). \tag{A.37}$$

Furthermore, solving eqn. (3.33) for $N_P(z)$ yields $N_P(z) = V(z)\, A(z)$. Inserting into eqn. (A.37) and dividing by $V(z)$ hence gives recursive relationship in the numerator of the transfer function as stated in eqn. (3.34) [107, ch. 8.7], [144, ch. 12.3.2].

A.4.4 Relation between reflection coefficients and acoustic tubes

This section shows that the acoustic tube can be represented by the same type of recursion as exhibited by the PARCOR coefficients as first mentioned in sect. §3.4.2.2 on page 55. Recalling the transfer function of the acoustic tube:

$$V(z) = \frac{\left\{ \frac{1}{2}\,(1 + r_G) \prod_{k=1}^{K} (1 + r_k) \right\} z^{-K/2}}{D(z)} \tag{3.5}$$

where the concatenated tubes are simplified to sections of equal lengths, $\Delta x = \ell/K$. The denominator is defined as

$$D(z) = \underbrace{\begin{bmatrix} 1 & -r_G \end{bmatrix} \underbrace{\begin{bmatrix} 1 & -r_1 \\ -r_1\, z^{-1} & z^{-1} \end{bmatrix}}_{\mathbf{P}_1} \cdots \begin{bmatrix} 1 & -r_{K+1} \\ -r_K\, z^{-1} & z^{-1} \end{bmatrix}}_{\mathbf{P}_K} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \mathbf{P}_K \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{3.6}$$

Assuming that $r_G = 1$ (unit gain) and solving for the first polynomial matrix, $\mathbf{P}_1$:

$$\begin{aligned}\mathbf{P}_1 &= \begin{bmatrix}1 & -r_G\end{bmatrix}\begin{bmatrix}1 & -r_1\\ -r_1 z^{-1} & z^{-1}\end{bmatrix} = \begin{bmatrix}1 + r_1 z^{-1} & -(r_1 + z^{-1})\end{bmatrix}\\ &= \begin{bmatrix}D_1(z) & -z^{-1} D_1(z^{-1})\end{bmatrix},\end{aligned} \tag{A.38}$$

where $D_1(z) = 1 + r_1 z^{-1} = D_0(z) + r_1 z^{-1} D_0(z^{-1})$ where $D_0(z) = 1$. Inserting eqn. (A.38) into $\mathbf{P}_2$:

$$\begin{aligned}\mathbf{P}_2 &= \mathbf{P_1}\begin{bmatrix}1 & -r_2\\ -r_2 z^{-1} & z^{-1}\end{bmatrix} = \begin{bmatrix}D_1(z) & -z^{-1} D_1(z^{-1})\end{bmatrix}\begin{bmatrix}1 & -r_2\\ -r_2 z^{-1} & z^{-1}\end{bmatrix}\\ &= \begin{bmatrix}D_1(z) + r_2 z^{-2} D_1(z^{-1}) & -r_2 D_1(z) - z^{-2} D_1(z^{-1})\end{bmatrix}\\ &= \begin{bmatrix}D_2(z) & -z^{-2} D_2(z^{-1})\end{bmatrix},\end{aligned} \tag{A.39}$$

where $D_2(z) = D_1(z) + r_2 z^{-2} D_1(z^{-1})$. Inserting into $\mathbf{P}_3$:

$$\begin{aligned}\mathbf{P}_3 &= \mathbf{P}_2\begin{bmatrix}1 & -r_3\\ -r_3 z^{-1} & z^{-1}\end{bmatrix} = \begin{bmatrix}D_2(z) & -z^{-2} D_2(z^{-1})\end{bmatrix}\begin{bmatrix}1 & -r_3\\ -r_3 z^{-1} & z^{-1}\end{bmatrix}\\ &= \begin{bmatrix}D_2(z) + r_3 z^{-3} D_2(z^{-1}) & -r_3 D_2(z) - z^{-3} D_2(z^{-1})\end{bmatrix}\\ &= \begin{bmatrix}D_3(z) & -z^{-3} D_3(z^{-1})\end{bmatrix},\end{aligned} \tag{A.40}$$

where $D_3(z) = D_2(z) + r_3 z^{-3} D_2(z^{-1})$. As, after this point, the structure of the matrices does not change anymore, it is safe to generalise the results in eqns. (A.39) and (A.40) such that the k^{th} polynomial matrix for some $1 \leq k \leq K$ can be written as

$$\mathbf{P}_k = \begin{bmatrix}D_k(z) & -z^{-k} D_k(z^{-1})\end{bmatrix} \tag{A.41}$$

where $D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1})$. Inserting for $k = K$ into eqn. (3.6):

$$D(z) = \mathbf{P}_k\begin{bmatrix}1\\ 0\end{bmatrix} = \begin{bmatrix}D_K(z) & -z^{-K} D_K(z^{-1})\end{bmatrix}\begin{bmatrix}1\\ 0\end{bmatrix} = D_K(z). \tag{A.42}$$

Thus, the denominator, $D(z)$, of the transfer function of the acoustic tube model can be expressed in terms of the recursions in eqn. (3.8).

A.5 Digital resonators

A.5.1 Relationship between source parameters and poles

This section derives the relationship between the parameters and poles of a second-order AR process as in sect. §3.4.3.4 on page 61. The system response, $H_{t,k}(z)$ of each digital resonator, $k \in \mathcal{K}$ can be obtained by taking the z-transform of the impulse response in eqn. (3.36), i.e.,

$$H_{t,k}(z) = \frac{g_{t,k}}{1 + a_{1,t,k}\, z^{-1} + a_{2,t,k}\, z^{-2}} \tag{A.43}$$

where $g_{t,k}$ is the resonant gain. Real poles correspond to pole pairs on the real axis [250]. Thus, as the phase of real poles is $\phi_{t,k} = 0$, eqn. (A.43) is equivalent to

$$H_{t,k}(z) = \frac{g_{t,k}}{1 + a_{1,t,k}\, z^{-1} + a_{2,t,k}\, z^{-2}} = \frac{g_{t,k}}{(1 - r_{1,t,k}\, z^{-1})(1 - r_{2,t,k}\, z^{-1})} \tag{A.44}$$

$$= \frac{g_{t,k}}{1 - (r_{1,t,k} + r_{2,t,k})z^{-1} + r_{1,t,k}\, r_{2,t,k}\, z^{-2}}. \tag{A.45}$$

where $r_{1,t,k}$ and $r_{2,t,k}$ are the locations of the poles on the real axis. Thus, for *real* pole pairs, the source parameters are related to the radii of the poles via

$$a_{1,t,k} = -(r_{1,t,k} + r_{2,t,k}) \qquad \text{and} \qquad a_{2,t,k} = r_{1,t,k}\, r_{2,t,k}. \tag{A.46}$$

As $-1 \leq \{r_{t,k},\, r^{\star}_{t,k}\} \leq 1$, the second resonator parameter maximum

$$a_{1,t,k} = \begin{cases} -2, & \text{if } r_{t,k} = r^{\star}_{t,k} = 1 \\ 2, & \text{if } r_{t,k} = r^{\star}_{t,k} = -1 \end{cases} \tag{A.47a}$$

$$a_{2,t,k} = \begin{cases} -1, & \text{if } r_{t,k} = \pm 1 \text{ and } r^{\star}_{t,k} = \mp 1 \\ 1, & \text{if } r_{t,k} = r^{\star}_{t,k} = 1 \end{cases} \tag{A.47b}$$

Therefore, the two resonator parameters are related in a triangular shape with the base between $-2 \leq a_{1,t,k} \leq 2$ and the height between $-1 \leq a_{2,t,k} \leq 1$ (see Fig. 3.15 on page 61). For *complex* pole pairs, the trigonometric identity

$$\cos \omega t = \frac{e^{\jmath \omega t} + e^{-\jmath \omega t}}{2} \tag{A.48}$$

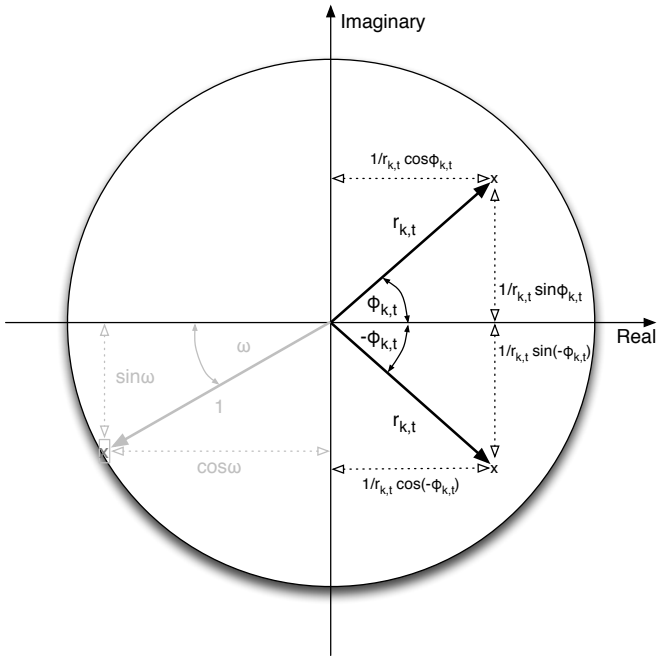


Figure A.1: *Poles of a digital resonator*

can be applied to eqn. (A.43), such that the system response is given as [251, 252]

$$H_{t,k}(z) = \frac{g_{t,k}}{(1 - p_{t,k}\, z^{-1})(1 - p^{\star}_{t,k}\, z^{-1})} \tag{A.49a}$$

$$= \frac{g_{t,k}}{(1 - r_{t,k}\, e^{\jmath\phi_{t,k}}\, z^{-1})\,(1 - r_{t,k}\, e^{-\jmath\phi_{t,k}}\, z^{-1})} \tag{A.49b}$$

$$= \frac{g_{t,k}}{1 - 2r_{t,k}\,\cos\phi_{t,k}\, z^{-1} + r^2_{t,k}\, z^{-2}}. \tag{A.49c}$$

where $g_{t,k}$ denotes the filter gain. Comparing eqns. (A.49a) and (A.49c), the parameters are related to the complex poles as stated in eqn. (3.45). Inserting the $0 \leq r_{t,k} \leq 1$ and $-1 \leq \cos\phi_{t,k} \leq 1$ corresponding to $0 \leq \phi_{t,k} \leq \pi$:

$$a_{1,t,k} = \begin{cases} -2 & \text{for } r_{t,k} = 1 \text{ and } \cos\phi_{t,k} = 1 \\ 2 & \text{for } r_{t,k} = 1 \text{ and } \cos\phi_{t,k} = -1 \end{cases} \tag{A.50}$$

Solving $a_{1,t,k}$ for $r_{t,k}$

$$r_{t,k} = \frac{-a_{1,t,k}}{2\cos\phi_{t,k}},$$

and inserting into $a_{2,t,k}$ yields

$$a_{2,t,k} = \frac{a^2_{1,t,k}}{4\cos^2\phi_{t,k}} = \begin{cases} 1 & \text{for } r_{t,k} = 1 \text{ and } \cos\phi_{t,k} = \pm 1 \\ 0 & \text{for } r_{t,k} = 0 \end{cases} \tag{A.51}$$

such that $0 \leq a_{2,t,k} \leq 1$. Recalling that $a_{1,t,k} = \pm 2$ for $\cos\phi_{t,k} = \mp 1$ according to eqn. (A.50), then by inserting these two conditions into eqn. (A.51), $a_{2,t,k}$ can be expressed as

$$a_{2,t,k} = \frac{1}{4} a_{1,t,k}^2, \tag{A.52}$$

which is a quadratic equation and thus $a_{2,t,k}$ varies with $a_{1,t,k}$ in the shape of a parabola with its minimum at 0 when $a_{1,t,k} = 0$ (Fig. 3.15). Inserting eqns. (A.50) and (A.52) for $a_{1,t,k}$ and $a_{2,t,k}$ in eqn. (3.36), the sub-signal of resonator $k \in \mathcal{K}$ is given as

$$x_{t,k} = 2r_{t,k} \cos\phi_{t,k}\, x_{k,t-1} - r_{t,k}^2\, x_{k,t-2} + g_{t,k}\, v_t \tag{A.53}$$

where $\sigma_{v_{t,k}} = g_{t,k}$ in eqn. (3.36). Eqn. (3.45) thus constitutes the relationship between the source parameters and the radii and phases of their corresponding phases. Using the remaining design specifications in eqn. (3.38) the resonant parameters can now be related to the poles and hence to the source parameters.

A.5.2 Relation of resonator frequency to pole phase

This section derives the relation of the resonant frequency with the complex poles as in sect. §3.4.3.2 on page 59. The magnitude response of a digital resonator can be found straightforwardly from eqn. (3.39) as

$$|H(\omega)| = \frac{g_{t,k}}{\sqrt{1 + r^2 - 2r\cos(\phi - \omega)}\sqrt{1 + r^2 - 2r\cos(\phi + \omega)}}. \tag{A.54}$$

As a resonance is a local peak in the magnitude response of the filter caused by a pole close to the unit circle, the resonant frequency is the frequency that maximises $|H(\omega)|$. As the gain, $g_{t,k}$ is independent of ω, the resonant frequency can thus be found by solving

$$\frac{\partial}{\partial\omega} \overbrace{\underbrace{\sqrt{1 + r^2 - \overbrace{2r\cos(\phi - \omega)}^{h(\omega)}}}_{u}\underbrace{\sqrt{1 + r^2 - 2r\cos(\phi + \omega)}}_{v}}^{g(h(\omega))}\Bigg|_{\omega=\omega_{t,k}} = 0. \tag{A.55}$$

where $\omega_{t,k}$ is the resonant frequency. Recalling the product rule of a derivative,

$$\frac{\partial uv}{\partial\omega} = \frac{\partial u}{\partial\omega}v + \frac{\partial v}{\partial\omega}u,$$

as well as the chain rule of a derivative,

$$\frac{\partial g\left(h(\omega)\right)}{\partial \omega} = \left(\frac{\partial g\left(h(\omega)\right)}{\partial h(\omega)}\right)\left(\frac{\partial h(\omega)}{\partial \omega}\right)$$

the left hand side (LHS) of the derivative is

$$\begin{aligned}\frac{\partial}{\partial \omega}\left(1+r^2-2r\,\cos\left(\phi\pm\omega\right)\right)^{1/2} &= \frac{1}{2u^{1/2}}\frac{\partial u}{\partial \omega}\\ &= \frac{-2r\left(-\sin\left(\phi\pm\omega\right)\right)\times(\pm 1)}{2u^{1/2}} = \frac{\mp r\,\sin\left(\phi\pm\omega\right)}{u^{1/2}}.\end{aligned}$$

Inserting for $\omega = \omega_{t,k}$, the derivative in eqn. (A.55) becomes

$$\begin{aligned}0 &= -\frac{r\sin\left(\phi-\omega_R\right)}{u^{1/2}}v^{1/2} + \frac{r\sin\left(\phi+\omega_R\right)}{v^{1/2}}u^{1/2}\\ &= -\frac{r\sin\left(\phi-\omega_R\right)v + r\sin\left(\phi+\omega_R\right)u}{(uv)^{1/2}}\\ &= -\frac{r\sin\left(\phi-\omega_R\right)v + r\sin\left(\phi+\omega_R\right)u}{(uv)^{1/2}}\\ &= \frac{r}{(uv)^{1/2}}\left[u\sin\left(\phi+\omega_R\right) - v\sin\left(\phi-\omega_R\right)\right]\end{aligned}$$

As $(uv)^{-1/2}$ cannot be zero, the derivative is equivalent to,

$$\begin{aligned}0 &= u\sin\left(\phi+\omega_R\right) - v\sin\left(\phi-\omega_R\right)\\ &= \left(1+r^2-2r\cos\left(\phi-\omega_R\right)\right)\sin\left(\phi+\omega_R\right)\\ &\quad - \left(1+r^2-2r\cos\left(\phi+\omega_R\right)\right)\sin\left(\phi-\omega_R\right)\\ &= \left(1+r^2\right)\sin\left(\phi+\omega_R\right) - 2r\cos\left(\phi-\omega_R\right)\sin\left(\phi+\omega_R\right)\\ &\quad - \left(1+r^2\right)\sin\left(\phi-\omega_R\right) - 2r\cos\left(\phi+\omega_R\right)\sin\left(\phi-\omega_R\right)\\ &= \left(1+r^2\right)\left[\sin\left(\phi+\omega_R\right) - \sin\left(\phi-\omega_R\right)\right]\\ &\quad - 2r\left[\cos\left(\phi-\omega_R\right)\sin\left(\phi+\omega_R\right) - \cos\left(\phi+\omega_R\right)\sin\left(\phi-\omega_R\right)\right],\end{aligned}$$

Reordering and applying the trigonometric identity:

$$\sin\left(A-B\right) = \sin A\cos B - \sin B\cos A \tag{A.56}$$

to the RHS yields

$$\left(1+r^2\right)\left[\sin\left(\phi+\omega_R\right) - \sin\left(\phi-\omega_R\right)\right] = 2r\sin 2\omega_R.$$

Applying the trigonometric identity

$$\sin A - \sin B = 2\cos\left(\frac{A+B}{2}\right)\sin\left(\frac{A-B}{2}\right) \tag{A.57}$$

to the LHS gives

$$2\left(1+r^2\right)\cos\phi\sin\omega_R = 2r\sin 2\omega_R. \tag{A.58}$$

Applying the trigonometric identity

$$\sin(A+B) = \sin A\cos B + \sin B\cos A \tag{A.59}$$

to the RHS:

$$\begin{aligned} 2\left(1+r^2\right)\cos\phi\sin\omega_R &= 2r\times 2\sin\omega_R\cos\omega_R, \\ \left(1+r^2\right)\cos\phi &= 2r\cos\omega_R, \end{aligned} \tag{A.60}$$

such that by solving for ω,

$$\omega_R = \cos^{-1}\left\{\frac{\left(1+r^2\right)\cos\phi}{2r}\right\}. \tag{A.61}$$

Thus, as the radial frequency is related to the frequency in cycles per sample via $\omega_R = 2\pi f$, the resonator frequency, $f_{t,k}$, is related to the phase of the pole, ϕ, via eqn. (3.40) on page 59.

A.5.3 Relation of resonator bandwidth to pole radius

This section derives the relation between the 3dB bandwidth of the digital resonator and the complex poles as in sect. §3.4.3.3 on page 59. The bandwidth of low-order filters is determined at an attenuated level - 3dB relevant to the maximum of the frequency response [148], i.e.,

$$\Delta\omega = \omega_2 - \omega_1, \tag{3.41}$$

where

$$|H(\omega_1)|^2 = |H(\omega_2)|^2 = \frac{|H(\omega_{t,k})|^2}{2}, \tag{A.62}$$

where $\omega_{t,k}$ is the resonant frequency,

$$\omega_{t,k} = \arccos\left\{\frac{(1+r^2)\cos\phi}{2r}\right\}. \tag{A.61}$$

The frequency response of a two-pole filter can be expressed by

$$\begin{aligned} H(\omega) &= \frac{g_{t,k}}{(1-p_1 e^{-\jmath\omega})(1-p_2 e^{-\jmath\omega})} = \frac{g_{t,k}}{(1-re^{\jmath\phi}e^{-\jmath\omega})(1-re^{-\jmath\phi}e^{-\jmath\omega})} \\ &= \frac{g_{t,k}}{1-2r\cos\phi e^{-\jmath\omega}+r^2e^{-2\jmath\omega}} \\ &= \frac{g_{t,k}}{1-2r\cos\phi\,(\cos\omega-\jmath\sin\omega)+r^2\,(\cos\omega-\jmath\sin\omega)} \\ &= \frac{g_{t,k}}{1-2r\cos\phi\cos\omega+r^2\cos 2\omega+\jmath\,(2r\cos\phi\sin\omega-r^2\sin 2\omega)}. \end{aligned}$$

The magnitude response can be obtained by multiplying the denominator with its complex conjugate,

$$|H(\omega)| = \frac{|g|}{\sqrt{\underbrace{\left(1-2r\cos\phi\cos\omega+r^2\cos 2\omega\right)^2+\left(2r\cos\phi\sin\omega-r^2\sin 2\omega\right)^2}_{D(\omega)^2}}} \tag{A.63}$$

such that the squared magnitude response is,

$$|H(\omega)|^2 = \frac{g_{t,k}^2}{D(\omega)^2}. \tag{A.64}$$

The squared denominator can be written as

$$\begin{aligned} D(\omega)^2 &= \left(1-2r\cos\phi\cos\omega+r^2\cos 2\omega\right)^2+\left(2r\cos\phi\sin\omega-r^2\sin 2\omega\right)^2 \\ &= 1+\boxed{4r^2\cos^2\phi\cos^2\omega}+\boxed{r^4\cos^2 2\omega}-4r\cos\phi\cos\omega \\ &\quad +2r^2\cos 2\omega\,\boxed{\boxed{-4r^3\cos\phi\cos\omega\cos 2\omega}} \\ &\quad +\boxed{4r^2\cos^2\phi\sin^2\omega}\,\boxed{\boxed{-4r^3\cos\phi\sin\omega\sin 2\omega}}+\boxed{r^4\sin^2 2\omega}. \end{aligned}$$

Noting that the terms in the oval and the squared box can be written respectively as

$$\begin{aligned} 4r^2\cos^2\phi\cos^2\omega+4r^2\cos^2\phi\sin^2\omega &= 4r^2\cos^2\phi\left(\cos^2\omega+\sin^2\omega\right), \\ r^4\cos^2 2\omega+r^4\sin^2 2\omega &= r^4\left(\cos^2 2\omega+\sin^2 2\omega\right), \end{aligned}$$

the trigonometric identity:

$$\cos^2 A + \sin^2 A = 1 \tag{A.65}$$

can be applied; Further noting that the terms in the double-framed box are equivalent to

$$\begin{aligned} &- 4r^3 \cos\phi \cos\omega \cos 2\omega - 4r^3 \cos\phi \sin\omega \sin 2\omega \\ &\qquad = -4r^3 \cos\phi \left(\cos\omega \cos 2\omega + \sin\omega \sin 2\omega\right), \end{aligned}$$

the trigonometric identity:

$$\cos\left(A - B\right) = \cos A \cos B + \sin A \sin B \tag{A.66}$$

can be applied; Thus, the squared denominator is equivalent to,

$$D(\omega)^2 = 1 - 4r\cos\phi\cos\omega + 2r^2\cos 2\omega + 4r^2\cos^2\phi - 4r^3\cos\phi\cos\omega + r^4.$$

Thus, the squared magnitude in eqn. (A.64) becomes,

$$|H(\omega)|^2 = \frac{g_{t,k}^2}{1 - 4r\cos\phi\cos\omega + 2r^2\cos 2\omega + 4r^2\cos^2\phi - 4r^3\cos\phi\cos\omega + r^4}. \tag{A.67}$$

To evaluate $|H(\omega_{t,k})|^2$, insert $\cos\omega_{t,k} = \frac{(1+r^2)}{2r}\cos\phi$ from eqn. (A.61) into $|H(\omega)|^2$, such that the denominator becomes,

$$\begin{aligned} D(\omega_{t,k})^2 = 1 &- 4r\cos\phi \times \frac{1+r^2}{2r}\cos\phi \\ &+ 2r^2\cos 2\omega + 4r^2\cos^2\phi - 4r^3\cos\phi \times \frac{1+r^2}{2r}\cos\phi + r^4. \end{aligned}$$

Applying the identity:

$$\cos^2 A = \frac{1}{2}\left(1 + \cos 2A\right) \tag{A.68}$$

to $4r^2\cos^2\phi$:

$$\begin{aligned}
D(\omega_{t,k})^2 &= 1 - 4r\cos\phi \times \frac{1+r^2}{2r}\cos\phi + 2r^2(2\cos^2\omega - 1) \\
&\quad + 4r^2\cos^2\phi - 4r^3\cos\phi \times \frac{1+r^2}{2r}\cos\phi + r^4 \\
&= 1 - 2(1+r^2)\cos^2\phi + 2r^2\left(2 \times \frac{(1+r^2)}{(2r)^2}\cos^2\phi - 1\right) \\
&\quad + 4r^2\cos^2\phi - 2r^2(1+r^2)\cos^2\phi + r^4 \\
&= 1 - 2(1+r^2)\cos^2\phi - 2r^2 + (1+r^2)^2\cos^2\phi + 4r^2\cos^2\phi \\
&\quad - 2r^2(1+r^2)\cos^2\phi + r^4 \\
&= \boxed{1} - 2\cos^2\phi - 2r^2\cos^2\phi \boxed{-2r^2} + \cos^2\phi + 2r^2\cos^2\phi + r^4\cos^2\phi \\
&\quad + 4r^2\cos^2\phi - 2r^2\cos^2\phi - 2r^4\cos^2\phi \boxed{+r^4}
\end{aligned}$$

rewriting the boxed terms as $1 - 2r^2 + r^4 = (1-r^2)^2$ and grouping together terms,

$$\begin{aligned}
&= (1-r^2)^2 - \cos^2\phi - r^4\cos^2\phi + 2r^2\cos^2\phi \\
&= (1-r^2)^2 - \cos^2\phi(1-r^2)^2 \\
&= (1-r^2)^2\left(1 - \cos^2\phi\right).
\end{aligned}$$

Thus, the squared magnitude response at resonant frequency is

$$|H(\omega_{t,k})|^2 = \frac{g_{t,k}^2}{(1-r^2)^2\,(1-\cos^2\phi)}. \tag{A.69}$$

Thus, in order to obtain $\omega_{1/2}$, equate $|H(\omega)|^2$ and $|H(\omega_{t,k})|^2/2$ and solve for ω, i.e.,

$$\frac{g_{t,k}^2}{2(1-r^2)^2\,(1-\cos^2\phi)} = \frac{g_{t,k}^2}{1 - 4r\cos\phi\cos\omega + 2r^2\cos 2\omega + 4r^2\cos^2\phi - 4r^3\cos\phi\cos\omega + r^4}$$

which is equivalent to

$$2(1-r^2)^2\left(1-\cos^2\phi\right) = 1\,\boxed{-4r\cos\phi\cos\omega} + 2r^2\cos 2\omega + 4r^2\cos^2\phi\,\boxed{-4r^3\cos\phi\cos\omega} + r^4.$$

Grouping in terms of $\cos\omega$,

$$\begin{aligned}
2(1-r^2)^2\left(1-\cos^2\phi\right) - 4r^2\cos^2\phi - r^4 - 1 &= -4r\cos\phi\cos\omega(1+r^2) + 2r^2\cos 2\omega \\
&= -4r\cos\phi\cos\omega(1+r^2) + 2r^2(2\cos^2\omega - 1).
\end{aligned}$$

Grouping in terms of terms independent of $\cos\omega$ on the LHS,

$$2(1-r^2)^2\left(1-\cos^2\phi\right)-4r^2\cos^2\phi-r^4-1+2r^2=-4r\cos\phi\cos\omega(1+r^2)+4r^2\cos^2\omega$$

Grouping in terms of $\cos^2\phi$ on the LHS,

$$\begin{aligned}
&-1+2(1-r^2)^2-r^4+2r^2-\cos^2\phi\left(2(1-r^2)^2+4r^2\right)\\
&\quad=-1+2(1-2r^2+r^4)-r^4+2r^2-2\cos^2\phi\left(1-2r^2+r^4+2r^2\right)\\
&\quad=(1-r^2)^2-2\cos^2\phi\left(1+r^4\right)
\end{aligned}$$

such that

$$\begin{aligned}
(1-r^2)^2-2\cos^2\phi\left(1+r^4\right)&=-4r\cos\phi\cos\omega(1+r^2)+4r^2\cos^2\omega\\
\frac{1}{4r^2}(1-4^2)^2-\frac{1}{2r^2}\cos^2\phi\left(1+r^4\right)&=-\frac{1}{r}\cos\phi\cos\omega(1+r^2)+\cos^2\omega
\end{aligned}$$

which leads to the quadratic equation,

$$\cos^2\omega-\cos\omega\underbrace{\cos\phi\frac{(1+r^2)}{r}}_{b}-\underbrace{\frac{1}{4r^2}(1-r^2)^2+\frac{1}{2r^2}\cos^2\phi\left(1+r^4\right)}_{c}=0$$

such that the solution of the quadratic formula is,

$$\begin{aligned}
\cos\omega_{1/2}&=\frac{-b\pm\sqrt{b^2-4c}}{2}\\
&=\frac{1}{2}\cos\phi\frac{(1+r^2)}{r}\\
&\quad\pm\frac{1}{2}\sqrt{\underbrace{\cos^2\phi\frac{(1+r^2)^2}{r^2}-4\times\left(-\frac{1}{4r^2}(1-r^2)^2+\frac{1}{2r^2}\cos^2\phi\left(1+r^4\right)\right)}_{S^2}}.
\end{aligned}$$

The term in the square root, S, can be written as

$$\begin{aligned}
S^2&=\cos^2\phi\frac{(1+r^2)^2}{r^2}+\frac{(1-r^2)^2}{r^2}-\frac{2}{r^2}\cos^2\phi(1+r^4)\\
&=\cos^2\phi\left(\frac{(1+r^2)^2}{r^2}-\frac{2(1+r^4)}{r^2}\right)+\frac{(1-r^2)^2}{r^2}=-\cos^2\phi\frac{(1-r^2)^2}{r^2}+\frac{(1-r^2)^2}{r^2}\\
&=\frac{(1-r^2)^2}{r^2}\left(1-\cos^2\phi\right)=\frac{(1-r^2)^2}{r^2}\sin^2\phi,
\end{aligned}$$

leading to eqn. (3.43) on page 60.

A.6 Room acoustical transfer function

This section derives the room acoustical transfer function in sect. §4.2 on page 66. The acoustic response in an enclosed space between a sound source and a receiver is the result of the direct-path signal and all its reflections. Thus the sound wave can be modelled by the superposition of all sound waves in the room. The sound propagation in terms of the sound pressure, $p(\mathbf{r}, t)$, can be described by the acoustic wave equation:

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c^2}\frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} = 0 \tag{4.1}$$

where $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ is the Laplacaian over the Cartesian coordinates $\mathbf{r} = (x, y, z)$. If a harmonic disturbance is producing the waves, for which the source function is given, then eqn. (4.1) can be rewritten as

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c^2}\frac{\partial^2 p(\mathbf{r}, t)}{\partial t^2} = -s(\mathbf{r}, t) \tag{A.70}$$

By application of the Fourier transform, the Helmholtz equation is obtained, i.e.,

$$\nabla^2 P(\mathbf{r}, \omega) + k^2 P(\mathbf{r}, \omega) = -S(\mathbf{r}, \omega), \tag{A.71}$$

where $P(\mathbf{r}, \omega)$ and $S(\mathbf{r}, \omega)$ are the Fourier transforms of $p(\mathbf{r}, t)$ and $s(\mathbf{r}, t)$ respectively, and k is $k = \omega/c$. For a unit-amplitude harmonic point source located at $\mathbf{r}_s = (x_s, y_s, z_s)$, the source function is $S(\mathbf{r}, \omega) = \delta(\mathbf{r} - \mathbf{r}_s)$. The partial differential equation in eqn. (A.71) thus can be solved by solving the inhomogeneous equation:

$$\nabla^2 H(\mathbf{r}, \mathbf{r}_s, \omega) + k^2 H(\mathbf{r}, \mathbf{r}_s, \omega) = -\delta(\mathbf{r} - \mathbf{r}_s) \tag{A.72}$$

where $H(\mathbf{r}, \mathbf{r}_s, \omega)$ is the RTF, describing the standing waves in the room. Note that as the RHS is a delta function and $H(\mathbf{r}, \mathbf{r}_s, \omega)$ is subject to a linear differential operator, the sound pressure for an arbitrary sound function, $S(\mathbf{r}, \omega)$ in eqn. (A.71) can be obtained by marginalising the source position, $\mathbf{r}_s$, from the product of the RTF and the sound source as shown in [4], i.e.,

$$P(\mathbf{r}, \omega) = \int H(\mathbf{r}, \mathbf{r}_s, \omega)\, S(\mathbf{r}_s, \omega) d\mathbf{r}_s \tag{A.73}$$

The solution of $p(\mathbf{r}, t)$ in eqn. (A.70) thus is the inverse Fourier transform of $P(\mathbf{r}, \omega)$.

In order to obtain $P(\mathbf{r}, \omega)$, an expression of the RTF, $H(\mathbf{r}, \mathbf{r}_s, \omega)$, in eqn. (A.72) required. $H(\mathbf{r}, \mathbf{r}_s, \omega)$ can be obtained by finding the eigenfunctions (or characteristic solutions), $P_i(\mathbf{r}, \omega)$, of the homogenous equations, $\nabla^2 H(\mathbf{r}, \mathbf{r}_s, \omega) + k^2 H(\mathbf{r}, \mathbf{r}_s, \omega) = 0$ [4].

In general, the RTF can thus be expressed of the form [4]

$$H(\mathbf{r}, \mathbf{r}_s, \omega) = \sum_{i=0}^{\infty} C_i(\mathbf{r}_s, \omega)\, P_i(\mathbf{r}, \omega) \tag{A.74}$$

where $C_i(\mathbf{r}_s, \omega)$ is a coefficient dependent on the source position, $\mathbf{r}_s$. The eigenfunctions, $P_i(\mathbf{r}, \omega)$, are mutually exclusive and satisfy

$$\int P_i(\mathbf{r}, \omega)\, P_j(\mathbf{r}, \omega) = \begin{cases} \alpha_i, & i = j \\ 0, & i \neq j \end{cases} \tag{A.75}$$

The general expression of the RTF in eqn. (A.74) can be made more physically meaningful by narrowing the specifications to rectangular rooms.

In a rectangular room with rigid walls, the eigenfunctions can be expressed as [4]:

$$P_i(\mathbf{r}, \omega) = P_i(\mathbf{r}) = \cos(k_x x)\cos(k_y y)\cos(k_z z) \tag{A.76}$$

where $k_\nu = {}^{(m_\nu \pi)}/_{L_\nu}$ with $\nu = \{x, y, z\}$ and where $m_\nu \in \mathbb{N}$ and the room is of dimension $L_x \times L_y \times L_z$ (width $\times$ length $\times$ height). By inserting eqn. (A.76) into eqn. (A.74), the RTF for rectangular rooms is determined as [4]

$$H(\mathbf{r}, \mathbf{r}_s, \omega) = \sum_{i=0}^{\infty} \frac{P_i(\mathbf{r})\, P_i(\mathbf{r}_s)}{\alpha_i (k^2 - k_i^2)} \tag{A.77}$$

where $k_i^2 = k_x^2 + k_y^2 + k_z^2 = {}^{\omega_i}/_{c}$ is the eigenvalue and ω_i is the i^{th} eigenfrequency at which the standing waves resonates (hence also known as resonant frequency). In realistic rooms with non-rigid walls, the eigenvalues, k_i^2 provide damping of resonance, such that k_i^2 should be expanded to $k_i^2 = {}^{\omega_i}/_{c} + \ell\delta_i/c$ where δ_i is the so-called damping constant, or Q-factor. If $\delta_i \ll \omega_i$, eqn. (A.77) can be rewritten as given by eqn. (4.3).

The inverse Fourier transform of the RTF, $H(\mathbf{r}, \mathbf{r}_s, \omega)$, gives the solution of the RIR, $h(\mathbf{r}, \mathbf{r}_s, t)$. As both $H(\mathbf{r}, \mathbf{r}_s, \omega)$ and $h(\mathbf{r}, \mathbf{r}_s, t)$ are expressed in terms of the source position, $\mathbf{r}_s$ and the receiving position, $\mathbf{r}_0$, both the RTF and RIR vary with changing source-sensor positions and distances [158, 159].

The expression of the RTF in eqn. (4.3) describes the acoustic properties of a reverberant room and can be used to model the reverberant channel. The following two sections discuss how the solution to the acoustic wave equation can be approximated by either simulating the impulse response of the room (sect. §4.3) or develop mathematical models based on the expression of the transfer function (sect. §4.4).

Appendix B

Background derivations: Methodology

B.1 MMSE estimators

This section derives the expression of the MMSE addressed first in sect. §5.2 on page 81. The MSE between the actual value of the desired variables, $\boldsymbol{\varphi}_{0:t}$, and their estimate, $\hat{\boldsymbol{\varphi}}_{0:t}$, can be expressed as

$$\begin{aligned}
\mathrm{MSE}_{\widehat{\mathbf{f}}_{0:t}} &= \mathbb{E}_{p(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)}\left[\|\widehat{\mathbf{f}}_{0:t}-\boldsymbol{\varphi}_{0:t}\|^2\right] = \int \|\widehat{\mathbf{f}}_{0:t}-\boldsymbol{\varphi}_{0:t}\|^2\; p\,(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)\,d\boldsymbol{\varphi}_{0:t} \\
&= \|\widehat{\mathbf{f}}_{0:t}\|^2 \int p\,(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)\,d\boldsymbol{\varphi}_{0:t} + \int \|\boldsymbol{\varphi}_{0:t}\|^2\; p\,(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)\,d\boldsymbol{\varphi}_{0:t} \\
&\qquad - 2\,\widehat{\mathbf{f}}_{0:t}^{\mathrm{T}} \int \boldsymbol{\varphi}_{0:t} p\,(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)\,d\boldsymbol{\varphi}_{0:t} \\
&= \|\widehat{\mathbf{f}}_{0:t}\|^2 + \mathbb{E}_{p(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)}\left[\|\boldsymbol{\varphi}_{0:t}\|^2\right] - 2\widehat{\mathbf{f}}_{0:t}^{\mathrm{T}}\,\mathbb{E}_{p(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)}\left[\boldsymbol{\varphi}_{0:t}\right]
\end{aligned} \tag{B.1}$$

where $\|\cdot\| \triangleq (\cdot)^{\mathrm{T}}(\cdot)$ is the Euclidean norm and

$$\mathbb{E}_{p(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)}\left[\boldsymbol{\varphi}_{0:t}\right] = \int \boldsymbol{\varphi}_{0:t}\, p\,(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)\,d\boldsymbol{\varphi}_{0:t} \tag{B.2}$$

is defined as the expected value. Differentiating eqn. (B.1) with respect to $\widehat{\mathbf{f}}_{0:t}$ and setting to zero in order to minimise the MSE yields

$$\frac{\partial \mathrm{MSE}_{\widehat{\mathbf{f}}_{0:t}}}{\partial \widehat{\mathbf{f}}_{0:t}} = 2\widehat{\mathbf{f}}_{0:t} - 2\mathbb{E}_{p(\,\boldsymbol{\varphi}_{0:t}\,|\,\mathbf{y}_{1:t}\,)}\left[\boldsymbol{\varphi}_{0:t}\right] = 0. \tag{B.3}$$

Solving for $\widehat{\mathbf{f}}_{0:t}$, the MMSE estimate can be expressed as stated eqn. (5.5) on page 82.

B.2 Optimality of the Kalman filter

This section shows that the Kalman filter is the optimal estimator in the MSE sense of the unknown states of conditionally Gaussian state-spaces (CGSSs) as first mentioned in sect. §5.3 on page 83. By definition, optimal estimators have estimation errors with zero mean and are orthogonal to the observations, $\mathbf{y}_{1:t}$. Therefore, in order to show that the Kalman filter is the optimal estimator, it is sufficient to show that the estimation errors have zero mean and are orthogonal to $\mathbf{y}_{1:t}$.

The error between the actual state and its *predicted* estimate is denoted as $\boldsymbol{\epsilon}_{t|t-1}$ and the error between $\mathbf{x}_t$ and its *updated* estimate is $\boldsymbol{\epsilon}_{t|t}$, where

$$\begin{aligned}\boldsymbol{\epsilon}_{t|t-1} &= \mathbf{x}_t - \boldsymbol{\mu}_{t|t-1} = \mathbf{A}_t\mathbf{x}_{t-1} + \mathbf{v}_t - \mathbf{A}_t\boldsymbol{\mu}_{t-1|t-1} = \mathbf{A}_t\left(\mathbf{x}_{t-1} - \boldsymbol{\mu}_{t-1|t-1}\right) + \mathbf{v}_t \\ &= \mathbf{A}_t\boldsymbol{\epsilon}_{t-1|t-1} + \mathbf{v}_t \qquad \text{(B.4)} \\ \boldsymbol{\epsilon}_{t|t} &= \mathbf{x}_t - \boldsymbol{\mu}_{t|t} = \mathbf{x}_t - \boldsymbol{\mu}_{t|t-1} - \mathbf{K}_t\left(\mathbf{y}_t - \mathbf{H}_t\boldsymbol{\mu}_{t|t-1}\right) \\ &= \boldsymbol{\epsilon}_{t|t-1} - \mathbf{K}_t\left(\mathbf{H}_t\mathbf{x}_t + \mathbf{w}_t - \mathbf{H}_t\boldsymbol{\mu}_{t|t-1}\right) = \left(\mathbf{I}_Q - \mathbf{K}_t\mathbf{H}_t\right)\boldsymbol{\epsilon}_{t|t-1} + \mathbf{K}_t\mathbf{w}_t \qquad \text{(B.5)}\end{aligned}$$

Assume that an initial optimal estimate is obtained at time $t = 1$, i.e.,

$$\mathbb{E}\left[\boldsymbol{\epsilon}_{1|1}\right] = \mathbf{0}_{Q\times 1} \qquad \text{and} \qquad \mathbb{E}\left[\boldsymbol{\epsilon}_{1|1}\mathbf{y}_1^T\right] = \mathbf{0}_{Q\times 1} \qquad \text{(B.6)}$$

where $\boldsymbol{\epsilon}_{t|t} \triangleq \mathbf{x}_t - \boldsymbol{\mu}_{t|t}$ is the error between the actual state of the system, $\mathbf{x}_t$, and its updated estimate, $\boldsymbol{\mu}_{t|t}$. By taking the expected value over the error terms, the mean of the error can be found as

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\epsilon}_{t|t-1}\right] &= \mathbf{A}_t\mathbb{E}\left[\boldsymbol{\epsilon}_{t-1|t-1}\right] - \mathbb{E}\left[\mathbf{v}_t\right] \\ \mathbb{E}\left[\boldsymbol{\epsilon}_{t|t}\right] &= \left(\mathbf{I}_Q - \mathbf{K}_t\mathbf{H}_t\right)\mathbb{E}\left[\boldsymbol{\epsilon}_{t|t-1}\right] + \mathbf{K}_t\mathbb{E}\left[\mathbf{v}_t\right]\end{aligned}$$

Thus, as both the excitation and measurement noise terms, $\mathbf{v}_t$ and $\mathbf{w}_t$, are zero-mean by definition, it can be shown by induction that the initial optimality in eqn. (B.6) is propagated forward in time such that $\mathbb{E}\left[\boldsymbol{\epsilon}_{t|t-1}\right] = \mathbb{E}\left[\boldsymbol{\epsilon}_{t|t}\right] = \mathbf{0}_{Q\times 1}$.

Orthogonality to the observations $\mathbf{y}_k$, $k \geq 1$ can be shown by taking the expected value over the linear product of the respective error terms and each observation, i.e.,

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\epsilon}_{t|t-1}\mathbf{y}_k^T\right] &= \mathbb{E}\left[\left(\mathbf{A}_t\boldsymbol{\epsilon}_{t-1|t-1} - \mathbf{v}_t\right)\mathbf{y}_k^T\right] = \mathbf{A}_t\mathbb{E}\left[\boldsymbol{\epsilon}_{t-1|t-1}\mathbf{y}_k^T\right] - \mathbb{E}\left[\mathbf{v}_t\mathbf{y}_k^T\right] \qquad \text{(B.7)} \\ \mathbb{E}\left[\boldsymbol{\epsilon}_{t|t}\mathbf{y}_k^T\right] &= \mathbb{E}\left[\left(\left(\mathbf{I}_Q - \mathbf{K}_t\mathbf{H}_t\right)\boldsymbol{\epsilon}_{t|t-1} + \mathbf{K}_t\mathbf{w}_t\right)\mathbf{y}_k^T\right] \\ &= \left(\mathbf{I}_Q - \mathbf{K}_t\mathbf{H}_t\right)\mathbb{E}\left[\boldsymbol{\epsilon}_{t|t-1}\mathbf{y}_k^T\right] + \mathbf{K}_t\mathbb{E}\left[\mathbf{w}_t\mathbf{y}_k^T\right] \qquad \text{(B.8)}\end{aligned}$$

Similar to the above argument, it can be shown by induction that $\mathbb{E}\left[\boldsymbol{\epsilon}_{t|t-1}\mathbf{y}_k^T\right] = \mathbb{E}\left[\boldsymbol{\epsilon}_{t|t}\mathbf{y}_k^T\right] = \mathbf{0}_{Q\times 1}$.

B.3 The MSE of the Kalman filter

This section shows that the diagonal terms in the covariance terms, $\boldsymbol{\mu}_{t|t-1}$ and $\boldsymbol{\mu}_{t|t}$, of the predicted and updated Kalman states correspond to the MSE of the corresponding states as first mentioned in sect. §5.3 on page 83. Recalling the definition of the error of the predicted and updated estimates in eqns. (B.4) and (B.5) in Appendix B.2, the covariance of the predicted and corrected estimates can be expressed as

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\epsilon}_{t|t-1}\boldsymbol{\epsilon}_{t|t-1}^T\right] &= \mathbb{E}\left[\left(\mathbf{A}_t\boldsymbol{\epsilon}_{t-1|t-1} + \mathbf{v}_t\right)\left(\mathbf{A}_t\boldsymbol{\epsilon}_{t-1|t-1} + \mathbf{v}_t\right)^T\right] \\
&= \mathbb{E}\left[\mathbf{A}_t\boldsymbol{\Sigma}_{t-1|t-1}\mathbf{A}_t^T + \boldsymbol{\Sigma}_{\mathbf{v}_t}\right] = \boldsymbol{\mu}_{t|t-1} \qquad \text{(B.9)} \\
\mathbb{E}\left[\boldsymbol{\epsilon}_{t|t}\boldsymbol{\epsilon}_{t|t}^T\right] &= \mathbb{E}\left[\left(\left(\mathbf{I}_Q - \mathbf{K}_t\mathbf{H}_t\right)\boldsymbol{\epsilon}_{t|t-1} + \mathbf{K}_t\mathbf{w}_t\right)\left(\left(\mathbf{I}_Q - \mathbf{K}_t\mathbf{H}_t\right)\boldsymbol{\epsilon}_{t|t-1} + \mathbf{K}_t\mathbf{w}_t\right)^T\right]
\end{aligned}$$

and after a little rearrangement and application of the Woodbury identity:

$$\left(\mathbf{A} + \mathbf{UCV}\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}$$

the expected becomes:

$$\mathbb{E}\left[\boldsymbol{\epsilon}_{t|t}\boldsymbol{\epsilon}_{t|t}^T\right] = \mathbb{E}\left[\left(\mathbf{I}_Q - \mathbf{K}_t\mathbf{H}_t\right)\boldsymbol{\Sigma}_{t|t-1}\right] = \boldsymbol{\mu}_{t|t}. \qquad \text{(B.10)}$$

Thus, the covariance of the error of the predicted and updated Kalman states is equivalent to the covariance of the predicted and updated Kalman states. The MSE a scalar estimate is defined as MSE $\triangleq \mathbb{E}\left[\epsilon^2\right]$ where ϵ is the scalar error. As $\boldsymbol{\varphi}_t \triangleq \begin{bmatrix} f_t & \dots & f_{t-Q+1} \end{bmatrix}^T$ the diagonal terms of the covariance matrix correspond to the MSEs of the corresponding elements in the predicted and updated states.

B.4 Optimal importance sampling function

This section derives the optimal importance sampling function and shows that it, indeed, minimises the variance of the MMSE estimator as mentioned in sect. §5.5.2 on page 96. The variance of the estimator can be expressed as

$$\operatorname{var}_{\pi(\boldsymbol{\varphi}_t|\mathbf{y}_{1:t},\boldsymbol{\varphi}_{0:t-1})}\left[w_t\right] = \mathbb{E}_{\pi(\boldsymbol{\varphi}_t|\mathbf{y}_{1:t},\boldsymbol{\varphi}_{0:t-1})}\left[w_t^2\right] - \left(\mathbb{E}_{\pi(\boldsymbol{\varphi}_t|\mathbf{y}_{1:t},\boldsymbol{\varphi}_{0:t-1})}\left[w_t\right]\right)^2. \qquad \text{(B.11)}$$

The term $\mathbb{E}_{\pi(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})}\left[w_t^2\right]$ can be obtained by inserting w_t as defined in eqn. (5.29) and solving the expected value, i.e.,

$$\begin{aligned}
\mathbb{E}_{\pi(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})}\left[w_t^2\right] &= \int w_t^2 \, \pi\left(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right) \mathrm{d}\boldsymbol{\varphi}_t \\
&= \int \left(w_{t-1} \frac{p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right)}{\pi\left(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right)} \right)^2 \pi\left(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right) \mathrm{d}\boldsymbol{\varphi}_t \\
&= w_{t-1}^2 \int \frac{\left(p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right)\right)^2}{\pi\left(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right)} \mathrm{d}\boldsymbol{\varphi}_t.
\end{aligned}$$

The term $\mathbb{E}_{\pi(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})}\left[w_t\right]$ is equivalent to

$$\begin{aligned}
&\mathbb{E}_{\pi(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1})}\left[w_t\right] \\
&\quad = \int w_{t-1} \frac{p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right)}{\pi\left(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right)} \pi\left(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right) \mathrm{d}\boldsymbol{\varphi}_t \\
&\quad = w_{t-1} \int p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right) \mathrm{d}\boldsymbol{\varphi}_t = p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}\right)
\end{aligned}$$

By inserting into eqn. (B.11), the variance thus is equivalent to eqn. (5.34) on page 96. The variance is zero if

$$\int \frac{\left(p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right)\right)^2}{\pi\left(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right)} \mathrm{d}\boldsymbol{\varphi}_t = p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}\right)^2 \tag{B.12}$$

If the proposal distribution is chosen according to eqn. (5.35) on page 96, then one can obtain by inserting eqn. (5.35) into eqn. (B.12):

$$\begin{aligned}
&\int \frac{\left(p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right)\right)^2}{\pi\left(\boldsymbol{\varphi}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\varphi}_{0:t-1}\right)} \mathrm{d}\boldsymbol{\varphi}_t \\
&\quad = \int \frac{\left(p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right)\right)^2}{p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right)} \times p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}\right) \mathrm{d}\boldsymbol{\varphi}_t \\
&\quad = \int p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right) p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}\right) \mathrm{d}\boldsymbol{\varphi}_t \\
&\quad = p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}\right) \times \int p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) p\left(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_{0:t-1}\right) \mathrm{d}\boldsymbol{\varphi}_t \\
&\quad = p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}\right) \times p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}\right) \\
&\quad = p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\varphi}_{0:t-1}\right)^2.
\end{aligned}$$

Thus, by inserting into eqn. (B.11) for the integral term, the two terms on the RHS cancel and thus the variance is 0.

Appendix C

Derivations of the RBPF

C.1 Augmented Kalman filter for source signal and channel estimation

This section derives expressions for the marginal estimators of the source signal and channel as referred to in sect. §6.4 on page 117.

Using eqns. (6.18a) and (6.18b) and assuming block-diagonality of the $\boldsymbol{\mu}_{P|P}$ and $\boldsymbol{\Sigma}_{P|P}$ the predicted states at time $P+1$ thus become:

$$\begin{aligned}
\boldsymbol{\mu}_{P+1|P} &= \begin{bmatrix} \mathbf{I}_{MP} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \mathbf{A}_{P+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{b},P} \\ \boldsymbol{\mu}_{\mathbf{x}_{P|P-1}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{b},P} \\ \boldsymbol{\mu}_{\mathbf{x}_{P|P-1}} \end{bmatrix} \\
\boldsymbol{\Sigma}_{P+1|P} &= \begin{bmatrix} \mathbf{0}_{MP\times MP} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \boldsymbol{\Sigma}_{\mathbf{v}_{P+1}} \boldsymbol{\Sigma}_{\mathbf{v}_{P+1}}{}^{\mathsf{T}} \end{bmatrix} \\
&\quad + \begin{bmatrix} \mathbf{I}_{MP} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \mathbf{A}_{P+1}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{b},P} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \boldsymbol{\Sigma}_{\mathbf{x}_{P|P}} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{MP} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \mathbf{A}_{P+1} \end{bmatrix} \\
&\triangleq \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{b},P} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}} \end{bmatrix}
\end{aligned} \tag{C.1}$$

where $\boldsymbol{\mu}_{\mathbf{x}_{P+1|P}} \triangleq \mathbf{A}_{P+1}\,\boldsymbol{\mu}_{\mathbf{x}_{P|P}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}} \triangleq \mathbf{A}_{P+1}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{x}_{P|P}} \mathbf{A}_{P+1} + \boldsymbol{\Sigma}_{\mathbf{v}_{P+1}} \boldsymbol{\Sigma}_{\mathbf{v}_{P+1}}{}^{\mathsf{T}}$. In other words, the block-diagonality of the initial states is retained by the Kalman prediction equations. Now, by application of eqns. (6.19a) and (6.19b), the residual covariance and Kalman gain are:

$$\begin{aligned}
\boldsymbol{\Sigma}_{z_{P+1}} &\triangleq \begin{bmatrix} \mathbf{Y}_P & \mathbf{C}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{b},P} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_P^{\mathsf{T}} \\ \mathbf{C} \end{bmatrix} + \boldsymbol{\Sigma}_{w_{P+1}}\,\boldsymbol{\Sigma}_{w_{P+1}}^{\mathsf{T}} \\
&= \mathbf{Y}_P\,\boldsymbol{\Sigma}_{\mathbf{b},P}\,\mathbf{Y}_P^{\mathsf{T}} + \mathbf{C}^{\mathsf{T}}\,\boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}}\,\mathbf{C} + \boldsymbol{\Sigma}_{w_{P+1}}\,\boldsymbol{\Sigma}_{w_{P+1}}^{\mathsf{T}}
\end{aligned} \tag{C.2}$$

which is a $M \times M$ matrix. Hence, the Kalman gain becomes:

$$\mathbf{K}_{P+1} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{b},P} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_P^T \\ \mathbf{C} \end{bmatrix} \boldsymbol{\Sigma}_{z_{P+1}}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{b},P}\,\mathbf{Y}_P^T\,\boldsymbol{\Sigma}_{z_{P+1}}^{-1} \\ \boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}}\,\mathbf{C}\,\boldsymbol{\Sigma}_{z_{P+1}}^{-1} \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{K}_{\mathbf{b}_{P+1}} \\ \mathbf{K}_{\mathbf{x}_{P+1}} \end{bmatrix} \tag{C.3}$$

where $\mathbf{K}_{\mathbf{b}_{P+1}} \triangleq \boldsymbol{\Sigma}_{\mathbf{b},P}\,\mathbf{Y}_P^T\,\boldsymbol{\Sigma}_{z_{P+1}}^{-1}$ and $\mathbf{K}_{\mathbf{x}_{P+1}} \triangleq \boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}}\,\mathbf{C}\,\boldsymbol{\Sigma}_{z_{P+1}}^{-1}$. By slightly reordering eqn. (6.18c), the updated states thus are expressed by:

$$\begin{aligned} \boldsymbol{\mu}_{P+1|P+1} &= \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{b},P} \\ \boldsymbol{\mu}_{\mathbf{x}_{P|P-1}} \end{bmatrix} + \begin{bmatrix} \mathbf{K}_{\mathbf{b}_{P+1}} \\ \mathbf{K}_{\mathbf{x}_{P+1}} \end{bmatrix} \left(\mathbf{y}_{P+1} - \begin{bmatrix} \mathbf{Y}_P & \mathbf{C}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{b},P} \\ \boldsymbol{\mu}_{\mathbf{x}_{P+1|P+1-1}} \end{bmatrix} \right) \\ &\triangleq \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{b},P+1} \\ \boldsymbol{\mu}_{\mathbf{x}_{P+1|P+1}} \end{bmatrix}. \end{aligned} \tag{C.4}$$

where $\boldsymbol{\mu}_{\mathbf{b},P+1} \triangleq \left(\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_{P+1}}\,\mathbf{Y}_P\right)\boldsymbol{\mu}_{\mathbf{b},P} + \mathbf{K}_{\mathbf{b}_{P+1}}\left(\mathbf{y}_{P+1} - \mathbf{C}^T\,\boldsymbol{\mu}_{\mathbf{x}_{P+1|P}}\right)$ and $\boldsymbol{\mu}_{\mathbf{x}_{P+1|P+1}} \triangleq \left(\mathbf{I}_Q - \mathbf{K}_{\mathbf{x}_{P+1}}\,\mathbf{Y}_P\right)\boldsymbol{\mu}_{\mathbf{x}_{P+1|P}} + \mathbf{K}_{\mathbf{x}_{P+1}}\left(\mathbf{y}_{P+1} - \mathbf{Y}_P\,\boldsymbol{\mu}_{\mathbf{b},P}\right)$. Similar to eqn. (C.4), by slightly reordering eqn. (6.18d), the updated covariance becomes:

$$\begin{aligned} \boldsymbol{\Sigma}_{P+1|P+1} &= \left(\mathbf{I}_{MP+Q} - \begin{bmatrix} \mathbf{K}_{\mathbf{b}_{P+1}} \\ \mathbf{K}_{\mathbf{x}_{P+1}} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_P & \mathbf{C}^T \end{bmatrix}\right) \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{b},P} & \mathbf{0}_{MP\times Q} \\ \mathbf{0}_{Q\times MP} & \boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}} \end{bmatrix} \\ &\triangleq \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{\mathbf{b},P+1} & \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}} \\ \hline \boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} & \boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P+1}} \end{array}\right] \end{aligned} \tag{C.5}$$

where the error covariance terms of the channel and source, $\boldsymbol{\Sigma}_{\mathbf{b},P+1}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P+1}}$ respectively, and the cross-correlations, $\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}$ and $\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}}$ are defined as

$$\boldsymbol{\Sigma}_{\mathbf{b},P+1} \triangleq \left(\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_{P+1}}\,\mathbf{Y}_P\right)\boldsymbol{\Sigma}_{\mathbf{b},P} \tag{C.6a}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P+1}} \triangleq \left(\mathbf{I}_Q - \mathbf{K}_{\mathbf{x}_{P+1}}\,\mathbf{C}^T\right)\boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}} \tag{C.6b}$$

$$\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}} \triangleq -\mathbf{K}_{\mathbf{b}_{P+1}}\,\mathbf{C}^T\,\boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}} \tag{C.6c}$$

$$\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} \triangleq -\mathbf{K}_{\mathbf{x}_{P+1}}\,\mathbf{Y}_P\,\boldsymbol{\Sigma}_{\mathbf{b},P}. \tag{C.6d}$$

By re-inserting into eqns. (6.18a) and (6.18b), the prediction at $P+2$ is expressed as:

$$\boldsymbol{\mu}_{P+2|P+1} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{b},P+1} \\ \mathbf{A}_{P+2}\,\boldsymbol{\mu}_{\mathbf{x}_{P+2|P+2}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{b},P+1} \\ \boldsymbol{\mu}_{\mathbf{x}_{P+2|P+2-1}} \end{bmatrix} \tag{C.7}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{P+2|P+1} &= \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{b},P+1} & \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\,\mathbf{A}_t \\ \mathbf{A}_t^T\,\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} & \mathbf{A}_{P+2}^T\boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P+1}}\mathbf{A}_{P+2} + \boldsymbol{\Sigma}_{\mathbf{v}_{P+2}}\,\boldsymbol{\Sigma}_{\mathbf{v}_{P+2}}{}^T \end{bmatrix} \qquad \text{(C.8)} \\ &\triangleq \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{b},P+1} & \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\,\mathbf{A}_t \\ \mathbf{A}_t^T\,\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} & \boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+1}} \end{bmatrix} \qquad \text{(C.9)} \end{aligned}$$

Therefore, the residual covariance at $P+2$ is:

$$\begin{aligned}\boldsymbol{\Sigma}_{z_{P+2}} &= \begin{bmatrix}\mathbf{Y}_P & \mathbf{C}^T\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{b},P+1} & \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_{P+2}\\ \mathbf{A}_{P+2}^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} & \boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+1}}\end{bmatrix}\begin{bmatrix}\mathbf{Y}_P^T\\ \mathbf{C}\end{bmatrix} + \boldsymbol{\Sigma}_{w_{P+1}}\boldsymbol{\Sigma}_{w_{P+1}}^T \\ &= \mathbf{Y}_P\boldsymbol{\Sigma}_{\mathbf{b},P}\mathbf{Y}_P^T + \mathbf{C}^T\boldsymbol{\Sigma}_{\mathbf{x}_{P+1|P}}\mathbf{C} + \boldsymbol{\Sigma}_{w_{P+1}}\boldsymbol{\Sigma}_{w_{P+1}}^T \\ &\quad + \mathbf{C}^T\mathbf{A}_{P+2}^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}}\mathbf{Y}_{t-1}^T + \mathbf{Y}_{t-1}\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_{P+2}\mathbf{C}\end{aligned} \tag{C.10}$$

i.e., due to the cross-correlation terms on the off-diagonals in $\boldsymbol{\Sigma}_{P+1|P+1}$, two additional terms $\mathbf{C}^T\mathbf{A}_{P+2}^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}}\mathbf{Y}_{t-1}^T$ and $\mathbf{Y}_{t-1}\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_{P+2}\mathbf{C}$ are introduced to the residual covariance as compared to at $P+1$. The Kalman gain is now expressed as

$$\mathbf{K}_{P+2} = \begin{bmatrix}\left(\boldsymbol{\Sigma}_{\mathbf{b},P+1}\mathbf{Y}_{P+1}^T + \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_t\mathbf{C}\right)\boldsymbol{\Sigma}_{z_{P+2}}\\ \left(\mathbf{A}_t^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}}\mathbf{Y}_P^T + \boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+1}}\mathbf{C}\right)\boldsymbol{\Sigma}_{z_{P+2}}\end{bmatrix} \triangleq \begin{bmatrix}\mathbf{K}_{\mathbf{b}_{P+2}}\\ \mathbf{K}_{\mathbf{x}_{P+2}}\end{bmatrix} \tag{C.11}$$

Again, additional terms due to the cross-correlations are now taken into account. The updated Kalman covariance can therefore be expressed as

$$\boldsymbol{\Sigma}_{P+2|P+2} = \left(\mathbf{I}_{MP+Q} - \begin{bmatrix}\mathbf{K}_{\mathbf{b}_{P+2}}\\ \mathbf{K}_{\mathbf{x}_{P+2}}\end{bmatrix}\begin{bmatrix}\mathbf{Y}_{P+1} & \mathbf{C}^T\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{b},P+1} & \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_{P+2}\\ \mathbf{A}_{P+2}^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} & \boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+1}}\end{bmatrix} \tag{C.12}$$

$$= \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{b},P+2} & \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+2}}\\ \boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+2}} & \boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+2}}\end{bmatrix} \tag{C.13}$$

where, by substituting $\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_t}$ and $\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_t}$ by eqns. (C.6d) and (C.6c) respectively:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{b},P+2} &= \boldsymbol{\Sigma}_{\mathbf{b},P+1} - \mathbf{K}_{\mathbf{b}_{P+2}}\mathbf{Y}_{P+1}\boldsymbol{\Sigma}_{\mathbf{b},P+1} - \mathbf{K}_{\mathbf{b}_{P+2}}\mathbf{C}^T\mathbf{A}_{P+2}^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} \\ &= \left(\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_{P+2}}\mathbf{Y}_{P+1}\right)\boldsymbol{\Sigma}_{\mathbf{b},P} - \mathbf{K}_{\mathbf{b}_{P+2}}\mathbf{C}^T\mathbf{A}_{P+2}^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}}\end{aligned} \tag{C.14}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+2}} &= \boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+1}} - \mathbf{K}_{\mathbf{x}_{P+2}}\mathbf{Y}_{P+1}\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_t - \mathbf{K}_{\mathbf{x}_{P+2}}\mathbf{C}^T\boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+1}} \\ &= \left(\mathbf{I}_Q - \mathbf{K}_{\mathbf{x}_{P+2}}\mathbf{C}^T\right)\boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+1}} - \mathbf{K}_{\mathbf{x}_{P+2}}\mathbf{Y}_{P+1}\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_{P+2}\end{aligned} \tag{C.15}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+2}} &= \boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_{P+2} - \mathbf{K}_{\mathbf{b}_{P+2}}\mathbf{Y}_{P+1}\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_{P+2} - \mathbf{K}_{\mathbf{b}_{P+2}}\mathbf{C}^T\boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+1}} \\ &= \left(\mathbf{I} - \mathbf{K}_{\mathbf{b}_{P+2}}\mathbf{Y}_{P+1}\right)\boldsymbol{\Sigma}_{(\mathbf{b}|\mathbf{x})_{P+1}}\mathbf{A}_{P+2} - \mathbf{K}_{\mathbf{b}_{P+2}}\mathbf{C}^T\boldsymbol{\Sigma}_{\mathbf{x}_{P+2|P+1}}\end{aligned} \tag{C.16}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+2}} &= \mathbf{A}_{P+2}^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} - \mathbf{K}_{\mathbf{x}_{P+2}}\mathbf{Y}_{P+1}\boldsymbol{\Sigma}_{\mathbf{b},P+1} - \mathbf{K}_{\mathbf{x}_{P+2}}\mathbf{C}^T\mathbf{A}_t^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} \\ &= \left(\mathbf{I} - \mathbf{K}_{\mathbf{x}_{P+2}}\mathbf{C}^T\right)\mathbf{A}_t^T\boldsymbol{\Sigma}_{(\mathbf{x}|\mathbf{b})_{P+1}} - \mathbf{K}_{\mathbf{x}_{P+2}}\mathbf{Y}_{P+1}\boldsymbol{\Sigma}_{\mathbf{b},P+1}\end{aligned} \tag{C.17}$$

As before, the updated states can be computed as:

$$\boldsymbol{\mu}_{P+2|P+2} = \begin{bmatrix} \left(\mathbf{I}_{MP} - \mathbf{K}_{\mathbf{b}_{P+2}} \boldsymbol{\Upsilon}_{P+1}\right) \boldsymbol{\mu}_{\mathbf{b},P+1} + \mathbf{K}_{\mathbf{b}_{P+2}} \left(\mathbf{y}_{P+2} - \mathbf{C}^{\mathsf{T}} \boldsymbol{\mu}_{\mathbf{x}_{P+2|P+2-1}}\right) \\ \left(\mathbf{I}_{Q} - \mathbf{K}_{\mathbf{x}_{P+2}} \mathbf{C}^{\mathsf{T}}\right) \boldsymbol{\mu}_{\mathbf{x}_{P+2|P+2-1}} + \mathbf{K}_{\mathbf{x}_{P+2}} \left(\mathbf{y}_{P+2} - \boldsymbol{\Upsilon}_{P+1} \boldsymbol{\mu}_{\mathbf{b},P+1}\right) \end{bmatrix} \tag{C.18}$$

$$\triangleq \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{b},P+2} \\ \boldsymbol{\mu}_{\mathbf{x}_{P+2|P+2}} \end{bmatrix}. \tag{C.19}$$

Hence, the block-separability of $\boldsymbol{\mu}_{t|t}$ and $\boldsymbol{\Sigma}_{t|t}$ is propagated in time due to the diagonality of $\mathbf{D}_t$ and $\boldsymbol{\Sigma}_{\mathbf{D}_t}$.

C.2 Marginalized likelihood function

This section derives expressions for the marginal likelihood as referred to in sect. §6.5 on page 120.

Recall from sect. §5.3 on page 83 that $p(\mathbf{z}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t})$ is the predicted posterior pdf parameterised by $\boldsymbol{\mu}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t-1}$ in eqns. (6.18a) and (6.18b). Using Bayes's theorem:

$$p(\mathbf{z}_t \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}_{0:t}) = \frac{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \mathbf{z}_t, \boldsymbol{\theta}_{0:t})\, p(\mathbf{z}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t})} \tag{C.20}$$

Then by inserting for the likelihood in eqn. (6.30) and the prediction pdf parameterised by eqns. (6.18a) and (6.18b) and integrating both sides with respect to zero, then the left hand side integrates to 1 and hence, by solving for $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t})$:

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}) = \frac{(2\pi)^{-MP+Q/2}}{|\boldsymbol{\Sigma}_{t|t-1}|^{\frac{1}{2}} |\boldsymbol{\Sigma}_{w_t} \boldsymbol{\Sigma}_{w_t}^{\mathsf{T}}|^{\frac{1}{2}}} \int_{\mathcal{Z}} \exp\Bigg\{ -\frac{1}{2} \Big[\mathbf{z}_t^{\mathsf{T}} \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{H}_t^{\mathsf{T}} (\boldsymbol{\Sigma}_{w_t} \boldsymbol{\Sigma}_{w_t}^{\mathsf{T}})^{-1} \mathbf{H}_t \right) \mathbf{z}_t$$
$$- 2\mathbf{z}_t^{\mathsf{T}} \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} \boldsymbol{\mu}_{t|t-1} + \mathbf{H}_t^{\mathsf{T}} (\boldsymbol{\Sigma}_{w_t} \boldsymbol{\Sigma}_{w_t}^{\mathsf{T}})^{-1} \mathbf{y}_t \right) + \boldsymbol{\mu}_{t|t-1}^{\mathsf{T}} \boldsymbol{\Sigma}_{t|t-1}^{-1} \boldsymbol{\mu}_{t|t-1} + \mathbf{y}_t^{\mathsf{T}} (\boldsymbol{\Sigma}_{w_t} \boldsymbol{\Sigma}_{w_t}^{\mathsf{T}})^{-1} \mathbf{y}_t \Big] \Bigg\} \mathrm{d}\mathbf{z}_t.$$

By application of the Gaussian identity:

$$\begin{aligned} p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}_{0:t}) &= \frac{(2\pi)^{-MP+Q/2}}{|\boldsymbol{\Sigma}_{t|t-1}|^{\frac{1}{2}} |\boldsymbol{\Sigma}_{w_t} \boldsymbol{\Sigma}_{w_t}^{\mathsf{T}}|^{\frac{1}{2}}} \frac{(2\pi)^{MP+Q/2}}{|\boldsymbol{\Sigma}_{t|t}|^{\frac{1}{2}}} \\ &\times \exp\Bigg\{ -\frac{1}{2} \Big[\boldsymbol{\mu}_{t|t-1}^{\mathsf{T}} \boldsymbol{\Sigma}_{t|t-1}^{-1} \boldsymbol{\mu}_{t|t-1} + \mathbf{y}_t^{\mathsf{T}} (\boldsymbol{\Sigma}_{w_t} \boldsymbol{\Sigma}_{w_t}^{\mathsf{T}})^{-1} \mathbf{y}_t - \boldsymbol{\mu}_{t|t}^{\mathsf{T}} \boldsymbol{\Sigma}_{t|t}^{-1} \boldsymbol{\mu}_{t|t} \Big] \Bigg\} \end{aligned} \tag{C.21}$$

Defining $\hat{\mathbf{K}}_t = \mathbf{I}_{MP+Q} - \mathbf{K}_t\mathbf{H}_t$, such that $\boldsymbol{\mu}_{t|t} = \hat{\mathbf{K}}_t\,\boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t\mathbf{y}_t$,

$$\begin{aligned}
&\boldsymbol{\mu}_{t|t-1}^T\boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{\mu}_{t|t-1} - \boldsymbol{\mu}_{t|t}^T\boldsymbol{\Sigma}_{t|t}^{-1}\boldsymbol{\mu}_{t|t} \\
&\quad= \boldsymbol{\mu}_{t|t-1}^T\left(\boldsymbol{\Sigma}_{t|t-1}^{-1} - \hat{\mathbf{K}}_t^T\boldsymbol{\Sigma}_{t|t}^{-1}\hat{\mathbf{K}}_t\right)\boldsymbol{\mu}_{t|t-1} - 2\mathbf{y}_t^T\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t}^{-1}\hat{\mathbf{K}}_t\boldsymbol{\mu}_{t|t-1} - \mathbf{y}_t^T\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t}^{-1}\mathbf{K}_t\mathbf{y}_t
\end{aligned}$$

Now, as $\boldsymbol{\Sigma}_{t|t} = \hat{\mathbf{K}}_t\boldsymbol{\Sigma}_{t|t-1}$,

$$= \boldsymbol{\mu}_{t|t-1}^T\left(\mathbf{I}_{MP+Q} - \hat{\mathbf{K}}_t^T\right)\boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{\mu}_{t|t-1} - 2\mathbf{y}_t^T\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{\mu}_{t|t-1} - \mathbf{y}_t^T\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t}^{-1}\mathbf{K}_t\mathbf{y}_t$$

and by inserting $\hat{\mathbf{K}}_t = \mathbf{I}_{MP+Q} - \mathbf{K}_t\mathbf{H}_t$:

$$= \boldsymbol{\mu}_{t|t-1}^T\mathbf{H}_t^T\,\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{\mu}_{t|t-1} - 2\mathbf{y}_t\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{\mu}_{t|t-1} - \mathbf{y}_t^T\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t}^{-1}\mathbf{K}_t\mathbf{y}_t$$

Using eqn. (6.19a), $\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^T\boldsymbol{\Sigma}_{z_t}^{-1}$ and hence $\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t-1}^{-1} = \boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{H}_t$, such that:

$$= \boldsymbol{\mu}_{t|t-1}^T\mathbf{H}_t^T\boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{H}_t\boldsymbol{\mu}_{t|t-1} - 2\mathbf{y}_t\boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{H}_t\boldsymbol{\mu}_{t|t-1} - \mathbf{y}_t^T\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t}^{-1}\mathbf{K}_t\mathbf{y}_t$$

Inserting into eqn. (C.21), the terms independent of $\boldsymbol{\mu}_{t|t-1}$ can be rewritten as

$$\begin{aligned}
&\mathbf{y}_t^T(\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1}\mathbf{y}_t - \mathbf{y}_t^T\mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t}^{-1}\mathbf{K}_t\mathbf{y}_t \\
&\quad= \mathbf{y}_t^T\left((\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1} - \mathbf{K}_t^T\boldsymbol{\Sigma}_{t|t}^{-1}\mathbf{K}_t\right)\mathbf{y}_t \\
&\quad= \mathbf{y}_t^T\left((\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1} - \mathbf{K}_t^T\left(\mathbf{H}_t^T(\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1}\mathbf{H}_t + \boldsymbol{\Sigma}_{t|t-1}^{-1}\right)\mathbf{K}_t\right)\mathbf{y}_t \\
&\quad= \mathbf{y}_t^T\left((\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1} - \mathbf{K}_t^T\mathbf{H}_t^T(\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1}\mathbf{H}_t\mathbf{K}_t - \boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{H}_t\mathbf{K}_t\right)\mathbf{y}_t
\end{aligned} \tag{C.22}$$

Inserting eqn. (6.19a) into eqn. (6.19b),

$$\boldsymbol{\Sigma}_{z_t} = \boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T + \mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^T = \boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T + \mathbf{H}_t\mathbf{K}_t\boldsymbol{\Sigma}_{z_t} \Rightarrow \mathbf{H}_t\mathbf{K}_t = \mathbf{I}_{MP} - \boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T\boldsymbol{\Sigma}_{z_t}^{-1} \tag{C.23}$$

such that eqn. (C.22) can be written as

$$\begin{aligned}
&\mathbf{y}_t^T\left((\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1} - \mathbf{K}_t^T\mathbf{H}_t^T(\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1}\left(\mathbf{I}_{MP} - \boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T\boldsymbol{\Sigma}_{z_t}^{-1}\right) - \boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{H}_t\mathbf{K}_t\right)\mathbf{y}_t \\
&\quad= \mathbf{y}_t^T\left((\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1} - \mathbf{K}_t^T\mathbf{H}_t^T(\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1}\right)\mathbf{y}_t \\
&\quad= \mathbf{y}_t^T\left(\mathbf{I}_{MP} - \mathbf{K}_t^T\mathbf{H}_t^T\right)(\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T)^{-1}\mathbf{y}_t = \mathbf{y}_t^T\boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{y}_t
\end{aligned}$$

Inserting into eqn. (C.21),

$$\begin{aligned} p\left(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta_{0:t}\right) = & \frac{(2\pi)^{-MP+Q/2}}{|\boldsymbol{\Sigma}_{t|t-1}|^{\frac{1}{2}}|\boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T|^{\frac{1}{2}}} \frac{(2\pi)^{MP+Q/2}}{|\boldsymbol{\Sigma}_{t|t}|^{\frac{1}{2}}} \\ & \times \exp\left\{-\frac{1}{2}\left[\mathbf{y}_t^T\boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{y}_t - 2\mathbf{y}_t^T\boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{H}_t\boldsymbol{\mu}_{t|t-1} + \boldsymbol{\mu}_{t|t-1}^T\mathbf{H}_t^T\boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{H}_t\boldsymbol{\mu}_{t|t-1}\right]\right\} \end{aligned} \tag{C.24}$$

Finally, note that the determinants in eqn. (C.24) are equivalent to

$$\frac{|\boldsymbol{\Sigma}_{t|t}|}{|\boldsymbol{\Sigma}_{t|t-1}|} = \det\left(\mathbf{I}_{MP} - \mathbf{K}_t\mathbf{H}_t\right) = \mathbf{I}_{MP} - \mathbf{H}_t\mathbf{K}_t = \boldsymbol{\Sigma}_{w_t}\boldsymbol{\Sigma}_{w_t}^T\boldsymbol{\Sigma}_{z_t}^{-1}$$

by using eqn. (C.23) and the identity $\det(\mathbf{I}_Q + \mathbf{u}\mathbf{v}^T) = 1 + \mathbf{v}^T\mathbf{y}$. Hence, eqn. (C.24) can be written in the simplified form as eqn. (6.32).

References

[1] C. Evers and J. R. Hopgood, "Parametric models for single-channel blind dereverberation of speech from a moving speaker," *IET J. Signal Process.*, vol. 2, no. 2, pp. 59–74, Jun. 2008.

[2] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[3] M. Bolić, P. M. Djurić, and S. Hong, "New resampling algorithms for particle filters," in *Proc. IEEE Conf. ICASSP*, 589-592, Ed., vol. 2, 2003.

[4] H. Kuttruff, *Room Acoustics*, 4th ed. London, England: Spon Press, 2000.

[5] A. K. Nábelek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1259–1265, 1989.

[6] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Royal Stat. Soc.*, vol. 21, pp. 577–580, 1949.

[7] B. Libbey and P. H. Rogers, "The effect of overlap-masking on binaural reverberant word intelligibility," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3141–3151, 2004.

[8] V. M. A. Peutz, "Articulation loss of consonants as a criterion for speech transmission in a room," *J. Audio Eng. Soc.*, vol. 19, no. 11, pp. 915–919, Dec. 1971.

[9] W. C. Knight, R. G. Pridham, and S. M. Kay, "Digital signal processing for sonar," *Proc. IEEE*, vol. 69, no. 11, pp. 1451–1508, Nov. 1981.

[10] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME - Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[11] A. Gelb, *Applied optimal estimation*. Cambridge, MA: MIT Press, May 1974.

[12] P. S. Maybeck, *Stochastic models, estimation, and control*. New York, NY: Academic Press, 1979, vol. 1.

[13] F. L. Lewis, *Optimal estimation with an introduction to stochastic control theory*. New Jersey, NJ: John Wiley & Sons, 1986.

[14] O. L. R. Jacobs, *Introduction to control theory*, 2nd ed. Oxford University Press, 1993.

[15] Y. Bar-Shalom and X.-R. Li, *Estimation and tracking: Principles, techniques and software*. YBS Publishing, 1998.

[16] A. Doucet, J. F. G. de Freitas, and N. J. Gordon, Eds., *Sequential Monte Carlo methods in practice*. New York, NY: Springer Verlag, 2000.

[17] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter - Particle Filters for Tracking Applications*. Artech House, 2004.

[18] J. V. Candy, *Bayesian signal processing: Classical, modern, and particle filtering meth-*

ods, ser. Technology & Engineering. New Jersey, NJ: John Wiley & Sons, 2009, vol. Volume 54 of Adaptive and learning systems for signal processing, communications, and control.

[19] B. D. O. Anderson and J. B. Moore, *Optimal filtering*. Englewood Cliffs, NJ: Prentice Hall, 1979.

[20] H. Tanizaki, *Nonlinear filters: estimation and applications*. New York, NY: Springer Verlag, 1996.

[21] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation*. New Jersey, NJ: John Wiley & Sons, 2001.

[22] J. M. Hammersley and K. W. Morton, "Poor man's Monte Carlo," *J. R. Statist. Soc.*, vol. 16, no. 23-38, 1954.

[23] N. J. Gordon, D. J. Salmond, and A. F. M. Salmond, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proceedings for Radar and Signal Processing*, vol. 140, no. 2, pp. 107–113, April 1993.

[24] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for bayesian filtering." *Statist. Comp.*, vol. 10, pp. 197–208, 2000.

[25] A. Doucet and X. Wang, "Monte Carlo methods for signal processing: A review in the statistical signal processing context," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 152–170, Nov. 2005.

[26] G. Casella and C. P. Robert, "Rao-Blackwellisation of sampling schemes," *Biometrika*, vol. 83, no. 1, pp. 81–94, 1996.

[27] G. Storvik, "Particle filters in state space models with the presence of unknown static parameters," *IEEE Trans. Signal Process.*, vol. 50, pp. 281–289, Feb. 2002.

[28] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, April 1988.

[29] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 3-4, pp. 229–240, Dec. 1996.

[30] N. D. Gaubitch and P. A. Naylor, "Analysis of the dereverberation performance of microphone arrays," in *Proc. IEEE Conf. IWAENC*, 2005, pp. 121–125.

[31] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 912–915, 1977.

[32] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1508–1515, Nov. 1985.

[33] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement usign beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[34] J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selective sound capture

for speech and audio processing," *Speech Commun.*, vol. 13, no. 1-2, pp. 207–222, Oct. 1993.

[35] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 425–437, 1997.

[36] N. Gaubitch, "Blind identification of acoustic systems and enhancement of reverberant speech," Ph.D. dissertation, Communications and Signal Processing Group, Department of Electrical and Electronic Engineering, Imperial College London, London, England, 2006.

[37] Y. Huang, J. Benesty, and J. Chen, "Dereverberation," in *Springer handbook of speech processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York, NY: Springer Verlag, 2008.

[38] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement usign cepstral processing," in *Proc. IEEE Conf. ICASSP*, vol. 2, 1991, pp. 977–980.

[39] M. Tohyama, R. H. Lyon, and T. Koike, "Source wave-form recovery in a reverberant space by cepstrum dereverberation," in *Proc. IEEE Conf. ICASSP*, vol. 1, 1993, pp. 157–160.

[40] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identificaton and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[41] R. Kennedy and B. Radlović, "Iterative cepstrum-based approach fro speech dereverberation," in *IEEE Proc. Int. Symp. Sig. Proc. App.*, vol. 1, 1999, pp. 55–58.

[42] A. Petropulu and S. Subramaniam, "Cepstrum based deconvolution for speech dereverberation," in *Proc. IEEE Conf. ICASSP*, vol. 1, 1994, pp. 1–12.

[43] S. Subramaniam, A. Petropulu, and C. Wendt, "Cepstrum based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 392–396, 1996.

[44] T. G. J. Stockham, T. M. Cannon, and R. B. Ingrbrestsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, no. 4, pp. 678–692, Apr. 1975.

[45] J. R. Hopgood, "Nonstationary signal processing with application to reverberation cancellation in acoustic environments," Ph.D. dissertation, University of Cambridge, Department of Engineering, Signal Processing Laboratory, Cambridge, UK, Nov. 2000.

[46] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[47] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models

for speech signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 6, pp. 1404–1413, Dec. 1985.

[48] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[49] J. Benesty, S. Makino, and J. Chen, Eds., *Speech enhancement*, ser. Signals and Commnication Technology. New York, NY: Springer Verlag, 2005.

[50] K. Lebart, J. M. Boucer, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, no. 3, pp. 359–366, 2001.

[51] S. F. Boll, "Suprression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 112–120, Apr. 1979.

[52] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. IEEE Conf. ICASSP*, vol. 4, Mar. 2005, pp. 173–176.

[53] J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. IEEE Conf. IWAENC*, Paris, France, 2006.

[54] E. A. P. Habets, S. Gannot, and I. Cohen, "Dereverberation and residual echo suppression in noisy environments," in *Speeech and audio processing in adverse environments*, ser. Signals and Commnication Technology, E. Hänsler and G. Schmidt, Eds. Berlin, Germany: Springer Verlag, 2008.

[55] M. Wu and D. Wang, "A two-stage algorithm for enhancement of reverberant speech," in *Proc. IEEE Conf. ICASSP*, vol. 1, 2005, pp. 1085–1088.

[56] B. W. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adpative filtering," in *Proc. IEEE Conf. ICASSP*, vol. 6, 2001, pp. 3701–3074.

[57] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. IEEE Conf. ICASSP*, Taipei, Taiwan 2009, pp. 3733–3736.

[58] M. S. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. IEEE Conf. ICASSP*, vol. 6, Seattle, WA, May 1998, pp. 3613–3616.

[59] H. Attias and L. Deng, "Speech denoising and dereverberation using probabilistic models," *Adv. Neural Info. Proc. Systems*, vol. 13, pp. 758–764, 2001.

[60] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, no. 2, pp. 175–184, Mar. 1952.

[61] T. Nakatani, M. Miyoshi, and K. Kinoshita, "Implementation and effects of single channel dereverberation based on the harmonic structure of speech," in *Proc. IEEE Conf. IWAENC*, 2003, pp. 91–94.

[62] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. IEEE Conf. ICASSP*, vol. 1, 2003, pp. 92–95.

[63] T. Nakatani, M. Miyoshi, and K. Kinoshita, "Single-microphone blind dereverberation," in *Speech enhancement*, ser. Signals and Commnication Technology, J. Benesty, S. Makino, and J. Chen, Eds. New York, NY: Springer Verlag, 2005, pp. 247–270.

[64] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

[65] T. Nakatani, B.-H. Juang, T. Hikichi, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Study on speech dereverberation with autocorrelation codebook," in *Proc. IEEE Conf. ICASSP*, vol. 1, Apr. 2007, pp. 193–196.

[66] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Harmonicity based dereverberation for improving automatic speech recognition performance and speech intelligibility," *IEICE Trans. Fund. Electr. Comm. and Comp. Sci.*, vol. E88-A, no. 7, pp. 1724–1731, 2005.

[67] E. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, Eindhoven, Netherlands, Jun. 2007.

[68] P. S. Spencer and P. J. W. Rayner, "Separation of stationary and time-varying systems and its application to the restoration of gramophone recordings," in *IEEE Proc. Int. Symp. Circ. Syst.*, vol. 1, Portland, OR, May 1989, pp. 292–295.

[69] P. S. Spencer, "System identification with application to the restoration of archived gramophone recordings," PhD Thesis, University of Cambridge, UK, Jun. 1990.

[70] C. Evers, "Single-channel blind dereverberation of speech from a moving speaker," MSc Thesis, The University of Edinburgh, UK, Aug. 2006.

[71] W. J. Hess, "Pitch and voicing determination," in *Advances in speech signal processing*, S. Furui and M. M. Sondhi, Eds. New York, NY: Marcel Dekker, 1992, pp. 3–48.

[72] M. S. Brandstein and S. M. Griebel, "Nonlinear, model-based microphone array speech enhancement," in *Acoustic signal processing for telecommunications*, S. L. Gay and J. Benesty, Eds. Boston, NY: Kluwer Academic Publishers, 2000, pp. 261–279.

[73] S. M. Griebel, "A microphone array system for speech source localization, denoising, and dereverberation," Ph.D. dissertation, Harvard University, Cambridge, MA, Apr. 2002.

[74] S. M. Griebel and M. S. Brandstein, "Microphone array speech dereverberation

usign coarse channel estimation," in *Proc. IEEE Conf. ICASSP*, vol. 1, 2001, pp. 201–204.

[75] S. M. Griebel and M. Brandstein, "Wavelet transform extrema clustering for multi-channel speech dereverberation," in *Proc. IEEE Conf. IWAENC*, Pocono Manor, PA, 1999, pp. 52–55.

[76] S. Griebel and M. Brandstein, "Wavelet transform extrema clustering for multi-channel speech dereverberation," in *Proc. IEEE Conf. WASPAA*, 1999, pp. 27–30.

[77] M. S. Brandstein and S. M. Griebel, "Nonlinear, model-based microphone array speech enhancement," *Acoustic Signal Processing for Telecommunication*, pp. 261–279, 2000.

[78] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Conf. ICASSP*, vol. 1, Orlando, FL, May 2002, pp. 541–544.

[79] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, 2000.

[80] N. Gaubitch, X. S. Lin, and P. A. Naylor, "Scale factor ambiguity correction for subband multichannel identification," in *Proc. IEEE Conf. IWAENC*, Seattle, WA, Sep. 2008.

[81] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *Int. Conf. Dig. Sig. Proc.*, 2007, pp. 607–610.

[82] N. D. Gaubitch, D. B. Ward, and P. A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4031–4039, Dec. 2006.

[83] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *Proc. EUSIPCO*, Vienna, Austria, Sep. 2004, pp. 809–812.

[84] N. D. Gaubitch, P. A. Naylor, and D. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. IEEE Conf. IWAENC*, Kyoto, Japan, 2003, pp. 99–102.

[85] M. Delcroix, "Speech dereverberation based on multi-channel linear prediction," Ph.D. dissertation, Graduate School of Information Science and Technology, Hokkaido University, Hokkaido, Japan, Mar. 2007.

[86] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denoising using multichannel linear prediction," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 6, pp. 1791–1801, Aug. 2007.

[87] ——, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[88] ——, "On the use of LIME dereverberation algorithm in an acoustic environ-

ment with a noise source," in *Proc. IEEE Conf. ICASSP*, vol. 1, 2006, pp. 825–828.

[89] ——, "Blind dereverberation algorithm for speech signals based on multi-channel linear prediction," *Acoustical Science and Technology*, vol. 26, no. 5, pp. 432–439, Jan. 2005.

[90] S. J. Godsill and C. Andrieu, "Bayesian separaton and recovery of convolutively mixed autoregressive sources," in *Proc. IEEE Conf. ICASSP*, vol. III, 1999, pp. 1733–1736.

[91] L. Deng, *Dynamic speech models: theory, algorithms, and applications*, ser. Synthesis lectures on speech and audio processing. Morgan & Claypool Publishers, 2006, vol. 2.

[92] F. Jelinek, *Statistical methods for speech recognition*. Cambridge, MA: MIT Press, 1997.

[93] K. Stevens, *Acoustic phonetics*. Cambridge, MA: MIT Press, 1998.

[94] D. Jurafsky and J. Martin, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Englewood Cliffs, NJ: Prentice-Hall, 2000.

[95] B. Gold and N. Morgan, *Speech and audio signal processing*. New Jersey, NJ: John Wiley & Sons, 2000.

[96] S. Furui, *Digital speech processing, synthesis and recognition*, 2nd ed. New York, NY: Marcel Dekker, Inc., 2001.

[97] X. D. Huang, A. Acero, and H. Hon, *Spoken language processing*. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[98] L. Deng and D. O'Shaughnessy, *Speech processing: A dynamic and optimization-oriented approach*. New York, NY: Marcel Dekker, Inc., 2003.

[99] B. Bloch and G. L. Trager, *Outline of linguistic analysis*. Baltimore, MD: Linguistic Society of America, 1942.

[100] J. L. Flanagan, *Speech analysis, synthesis and perception*, 2nd ed. New York, NY: Springer Verlag, 1972.

[101] K. Honda, "Physiological processes of speech production," in *Springer handbook of speech processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York, NY: Springer Verlag, 2008.

[102] R. Linggard, *Electronic synthesis of speech*. Cambridge, UK: Cambridge University Press, 1985.

[103] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 106, pp. 1511–1522, 1999.

[104] K. Ishizaka, M. Matsudaira, and T. Kaneko, "Input acoustic-impedance measurement of the subglottal system," *J. Acoust. Soc. Amer.*, vol. 60, pp. 190–197, 1976.

[105] R. J. Baken and R. F. Orlikoff, *Clinical measurement of speech and voice*. Singular Publishing Group, Aug. 1996.

[106] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Reading, MA: Addison-Wesley, 1987.

[107] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[108] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proc. Int. Congress Acoustics*, Copenhagen, Denmark, 1962, pp. 1–4.

[109] J. D. Markel and A. H. Gray, *Linear prediction of speech*. New York, NY: Springer Verlag, 1976.

[110] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of acoustics*, 4th ed. New Jersey, NJ: John Wiley & Sons, 2000.

[111] T. Beierholm and O. Winther, "Particle filter inference in an articulatory-based speech model," *IEEE Sig. Process. Lett.*, vol. 14, no. 11, pp. 883–886, Nov. 2007.

[112] J. L. Flanagan, "Note on the design of "terminal-analog" speech synthesizers," *J. Acoust. Soc. Amer.*, vol. 29, no. 2, pp. 306–310, Feb. 1957.

[113] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 67, no. 3, pp. 971–995, Mar. 1980.

[114] J. V. Candy, *Model-based signal processing*. New Jersey, NJ: John Wiley & Sons, 2006.

[115] H. D. Tran, K. Takeda, and F. Itakura, "A speech enhancement system based on data clustering and cumulative histogram equalization," in *IEEE Conf. Data Eng. Workshops*, Apr. 2005, p. 1207.

[116] I. J. Gdoura, P. Loizou, and A. Spanias, "Speech processing using higher order statistics," in *Proc. IEEE Conf. ISCAS*, May 1993, pp. 160–163.

[117] D. Aboutajdine, A. Adib, and A. Meziane, "Fast adpative algorithms for AR parameter estimation using higher order statistics," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 1998–2009, Aug. 1996.

[118] E. Nemer, R. Goubran, and S. Mahmoud, "Speech enhancement using fourth-order cumulants and optimum filters in the subband domain," *Speech Communications*, vol. 36, no. 3-4, pp. 219–246, Mar. 2002.

[119] J. Vermaak, M. Niranjan, and S. J. Godsill, "An improved speech production model for voiced speech utilising a seasonal AR-AR model and Markov chain Monte Carlo simulation," Cambridge University, UK, CUED/F-INFENG/TR.325, June 1998.

[120] J. Makhoul, "Linear prediction: a tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–256, August 1975.

[121] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*. Englewood Cliffs, NJ: Macmillan Publishing Company, 1993.

[122] A. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, no. 2, pp. 583–588, 1971.

[123] P. C. Loizou, *Speech enhancement: Theory and practice*. CRC Press, 2007.

[124] T. Quatieri, *Discrete-time speech signal processing*. Englewood Cliffs, NJ: Prentice-Hall, 2002.

[125] H. Teager and S. Teager, "Evidence for non-linear sound production mechanisms in the vocal tract," in *Speech production and speech modelling*, W. Hardcastle and A. Marchal, Eds. Boston, MA: Kluwer Academic Publishers, 1990, pp. 241–262.

[126] E. Bognar and H. Fujisaki, "Analysis, synthesis and perception of the French nasal vowels," in *Proc. IEEE Conf. ICASSP*, vol. 11, Apr. 1986, pp. 1601–1604.

[127] S. M. Kay, *Fundamentals of Statistical Processing, Volume I: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[128] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," *Signal Proc.*, vol. 5, no. 3, pp. 267–285, May 1978.

[129] T. S. Rao, "The fitting of nonstationary time-series models with time-dependent parameters," *J. Royal Stat. Soc. B*, vol. 32, no. 2, pp. 312–322, 1970.

[130] P. Gruber and J. Todtli, "Estimation of quasiperiodic signal parameters by means of dynamic signal models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 42, no. 3, pp. 552–562, Mar. 1994.

[131] A. Doucet, S. Godsill, and M. West, "Monte Carlo filtering and smoothing with application to time-varying spectral estimation," in *Proc. IEEE Conf. ICASSP*, Istanbul, Turkey, Jun. 2000, pp. 701–704.

[132] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 173–185, Mar. 2002.

[133] K. M. Malladi and R. V. Rajakumar, "Estimation of time-varying AR models of speech through Gauss-Markov modeling," in *Proc. IEEE Conf. ICASSP*, vol. 6, April 2003, pp. 305–308.

[134] H. M. James, N. B. Nichols, and R. S. Phillips, *Theory of servomechanisms*. New York, NY: McGraw-Hill, 1946.

[135] G. E. Dullerud and S. G. Lall, "Analysis and synthesis tools for time-varying systems," in *Proc. IEEE Conf. Dec. & Control*, vol. 4, no. 30, San Diego, CA, Dec. 1997, pp. 4543–4548.

[136] A. M. Lyapunov, *Stability of motion*. New York, NY: Academic Press, 1966.

[137] B. D. O. Anderson and J. B. Moore, "New results in linear system stability," *SIAM J. Control*, vol. 7, no. 3, pp. 398–414, Aug. 1969.

[138] R. Kalman, "On the stability of time-varying linear systems," *IEEE Trans. Circuit Theory*, vol. 9, no. 4, pp. 420–422, Dec. 1962.

[139] M. Juntunen, J. Tervo, and J. P. Kaipio, "Stabilization of stationary and time-varying autoregressive models," in *Proc. IEEE Conf. ICASSP*, 1998, pp. 2173–2176.

[140] ——, "Stabilization of Subba Rao-Liporace models," *Circuits Syst. Signal Process.*, vol. 18, pp. 395–406, 1999.

[141] M. Juntunen and J. P. Kaipio, "Stabilization of smoothness priors time-varying autoregressive models," *Circuits Syst. Signal Process*, vol. 19, no. 5, pp. 423–435, 2000.

[142] C. W. Therrien, *Discrete random signals and statistical signal processing*, ser. Signal processing series. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[143] W. Fong, S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing with application to audio signal enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 438–449, Feb. 2002.

[144] J. G. Proakis and D. G. Manolakis, *Digital signal processing*, 4th ed. Englewood Cliffs, NJ: Prentice Hall, 2007.

[145] S. Haykin, *Adaptive filter theory*, ser. Information and System Science Series, T. Kailath, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[146] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: Forecasting and control*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[147] T. van Waterschoot and M. Moonen, "A pole-zero placement technique for designing second-order IIR parametric equalizer filters," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 8, pp. 2561–2565, Nov. 2007.

[148] M. Cherniakov, *An introduction to parametric digital filters and oscillators*. New Jersey, NJ: John Wiley & Sons, 2003.

[149] E. C. Cherry and W. K. Taylor, "Some further experiments upon recognition of speech with one and with two ears," *J. Royal Stat. Soc.*, vol. 26, pp. 554–559, 1954.

[150] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 1633–1654, October 1999.

[151] A. J. Watkins and N. J. Holt, "Effects of a complex reflection on vowel identification," *Acoustica*, vol. 86, pp. 532–542, 2000.

[152] Y. Takata and A. K. Nábelek, "English consonant recognition in noise and in reverberation by Japanese and American listeners," *J. Acoust. Soc. Amer.*, vol. 88, pp. 663–666, 1990.

[153] A. K. Nábelek and D. Mason, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *Journal of Speech and Hearing Research*, vol. 24, pp. 375–383, 1981.

[154] J. R. Hopgood, "Audio signal processing in acoustic environments: Modelling reverberation and dereverberation," in *Tutorial, EUSIPCO*, Antalya, Turkey, 2005.

[155] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[156] M. Kompis and N. Dillier, "Simulating transfer functions in a revererbant room including source directivity and head-shadow effects," *J. Acoust. Soc. Amer.*, vol. 93, no. 5, pp. 2779–2787, May 1993.

[157] C. Thompson, K. Chandra, and V. Mehta, "Simulation of teleconferencing environments," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New York, October 1995, pp. 131–136.

[158] D. Cole, M. Moody, and S. Sridharan, "Position-independent enhancement of reverberant speech," *J. Audio Eng. Soc.*, vol. 45, no. 3, pp. 142–147, 1997.

[159] J. N. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *J. Sound Vibr.*, vol. 102, no. 2, pp. 217–228, September 1985.

[160] J. N. Mourjopoulos and M. A. Paraskevas, "Pole and zero modeling of room transfer functions," *J. Sound Vibr.*, vol. 146, no. 2, pp. 281–302, April 1991.

[161] J. Mourjopoulos and J. K. Hammond, "Modelling and enhancement of reverberant speech using an envelope convolution method," in *Proc. IEEE Conf. ICASSP*, 1983, pp. 1144–1147.

[162] Y. Haneda, Y. Kaneda, and N. Kitawaki, "Common-acoustical-pole and residue model and its application to spatial interpolation and extrapolation of a room transfer function," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 709–717, Nov. 1999.

[163] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 320–328, Apr. 1994.

[164] H. Wang and F. Itakura, "An implementation of multi-microphone dereverberation appraoch as a preprocessor to the word recognition system," *J. Acoust. Soc. Japan*, vol. 13, no. 5, pp. 285–293, 1992.

[165] ——, "Dereverberation of speech signals based on sub-band envelope estimation," *IEICE Trans. Fund. Electr. Comm. and Comp. Sci.*, vol. E74-A, no. 11, pp. 3576–3583, 1991.

[166] J. R. Hopgood, "A subband modelling approach to the enhancement of speech captured in reverberant acoustic environments: MIMO case," in *Proc. IEEE Conf. WASPAA*, Mohonk, Oct. 2005.

[167] T. Hidaka, L. L. Beranek, S. Masuda, N. Nishihara, and T. Okano, "Acoustical design of the Tokyo Opera City (TOC) concert hall, Japan," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 340–354, Jan. 2000.

[168] E. Demidenko, *Mix models: theory and applications*, ser. Wiley Series in probability and mathematical statistics. Probability and mathematical statistics. Wiley-

IEEE, 2004, vol. 518.

[169] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[170] R. Pintelon and J. Schoukens, *System identification: a frequency domain approach*. New Jersey, NJ: John Wiley & Sons, 2001.

[171] A. M. Johansen, "Some non-standard sequential Monte Carlo methods their applications," Ph.D. dissertation, University of Cambridge, Signal Processing Laboratory, Cambridge, UK, Dec. 2006.

[172] S. Gannot and A. Yeredor, "The Kalman filter," in *Springer handbook of speech processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York, NY: Springer Verlag, 2008.

[173] S. F. Schmidt, "Computational techniques in kalman filtering," NATO Advisory Group for Aerospace Research and Development, London, England, Tech. Rep., 1970.

[174] ——, "Practical aspects of kalman filtering implementation," NATO Advisory Group for Aerospace Research and Development, London, England, Tech. Rep., 1976.

[175] R. van der Merwe, "Sigma-point Kalman filters for probabilistic inference in dynamic state-space models," Ph.D. dissertation, OGI School of Science & Engineering, Oregon Health & Science University, Apr. 2004.

[176] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[177] E. A. Wan and R. van der Merwe, *Kalman filtering and neural networks*, ser. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control, S. Haykin, Ed. New Jersey, NJ: John Wiley & Sons, 2001.

[178] R. van der Merwe, A. Doucet, N. de Freitas, and E. A. Wan, "The unscented particle filter," Cambridge University, Engineering Department, CUED/F-INFENG/TR.380, Aug. 2000.

[179] E. A. Wan, R. van der Merwe, and A. T. Nelson, "Dual estimation and the unscented transformation," *Adv. Neural Info. Proc. Systems*, vol. 12, pp. 666–672, 2000.

[180] A. F. M. Smith and G. O. Roberts, "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *J. R. Statist. Soc.*, vol. 55, no. B, pp. 3–24, 1993.

[181] B. Y. Rubinstein, *Simulation and the Monte Carlo method*. New Jersey, NJ: John Wiley & Sons, 1981.

[182] J. Geweke, "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica*, vol. 57, pp. 1317–1339, 1989.

[183] M. Isard and A. Blake, "Condensation - conditional density propagation for vi-

sual tracking," *Int. J. Comp. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.

[184] K. Kanazawa, D. Koller, and S. Russell, "Stochastic simulation algorithms for dynamic probabilistic networks," in *Proc. Conf. Uncertainty in Artificial Intelligence*, 1995, pp. 346–351.

[185] P. Del Moral, "Measure-valued processes and interacting particle systems. Application to nonlinear filtering problems," *Ann. Appl. Probab.*, vol. 8, no. 2, pp. 438–495, 1998.

[186] V. S. Zaritskii, V. B. Svetnik, and L. I. Shimelevich, "Monte Carlo technique in problems of optimal data processing," *Automation and Remote Control*, vol. 12, pp. 95–103, 1975.

[187] H. Akashi and H. Kumamoto, "Random sampling approach to state estimation in switching environments," *Automatica*, vol. 13, pp. 429–434, 1977.

[188] R. Chen and J. S. Liu, "Predictive updating methods with application to Bayesian classification," *J. Royal Stat. Soc. B*, vol. 58, pp. 397–415, 1996.

[189] A. Kong, J. S. Liu, and W. H. Wong, "Sequential imputations and Bayesian missing data problems," *J. Am. Stat. Assoc.*, vol. 89, pp. 278–288, 1994.

[190] J. S. Liu and R. Chen, "Blind deconvolution via sequential imputation," *J. Am. Stat. Assoc.*, vol. 90, pp. 567–576, 1995.

[191] A. Doucet, "Algorithmes Monte Carlo pour l'estimation bay/'esienne de modèles markoviens cachés. Application au traitement de signaux de rayonnements," Ph.D. dissertation, Université de Paris-Sud Orsay, 1997.

[192] J. E. Handschin, "Monte Carlo techniques for prediction and filtering of non-linear stochastic processes," *Automatica*, vol. 6, pp. 555–563, 1970.

[193] J. E. Handschin and D. Q. Mayne, "Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering," *Intern. J. Control*, vol. 9, pp. 547–559, 1969.

[194] H. Tanizaki and R. S. Mariano, "Nonlinear and non-Gaussian state-space modeling with Monte-Carlo simulations," *J. Econometrics*, vol. 83, pp. 263–290, 1998.

[195] J. D. Hol, T. B. Schon, and F. Gustaffson, "On resampling algorithms for particle filters," in *IEEE Proc. Nonlinear Stat. Sig. Proc. Workshop*, 2006, pp. 79–82.

[196] R. Douc, O. Cappe, and E. Moulines, "Comparison of resampling schemes for particle filtering," in *Proc. IEEE Conf. ISPA*, 2005, pp. 64–69.

[197] T. Higuchi, "Self organizing time series model," in *Sequential Monte Carlo methods in practice*, A. Doucet, J. F. G. de Freitas, and N. J. Gordon, Eds. New York: Springer Verlag, 2000, pp. 429–444.

[198] C. P. Robert and G. Casella, *Monte Carlo statistical methods*, ser. Springer texts in statistics. New York, NY: Springer Verlag, 2004.

[199] J. S. Liu and R. Chen, "Sequential Monte Carlo methods for dynamic systems," *J. Am. Stat. Assoc.*, vol. 93, no. 443, pp. 1032–1044, September 1998.

[200] S. Godsill, A. Doucet, and M. West, "Maximum a posteriori sequence estimation using Monte Carlo particle filters," *Ann. Inst. Statist. Math.*, vol. 53, no. 1, pp. 82–96, March 2001.

[201] W. R. Gilks and C. Berzuini, "Following a moving target – Monte Carlo inference for dynamic Bayesian models," *J. Royal Stat. Soc. B*, vol. 63, pp. 127–146, 2001.

[202] J. S. Liu, "Metropolized independent sampling with comparison to rejection sampling and importance sampling," *Statistics and Computing*, vol. 6, no. 113-119, 1996.

[203] S. N. MacEachern, M. Clyde, and J. S. Liu, "Sequential importance sampling for nonparametric Bayes models: The next generation," *Can. J. Statist.*, vol. 27, no. 2, pp. 251–267, June 1999.

[204] P. Li, R. Goodall, and V. Kadirkamanathan, "Estimation of parameters in a linear state space model using a rao-blackwellised particle filter," *IET Proc.-Control Theory Appl.*, vol. 151, no. 6, pp. 727–738, November 2004.

[205] G. Hendeby, R. Karlsson, and F. Gustafsson, "A new formulation of the Rao-Blackwellized particle filter," in *IEEE Proc. Workshop Stat. Sig. Proc.*, Aug. 2007, pp. 84–88.

[206] T. Schön, F. Gustafsson, and P.-J. Nordlund, "Marginalized particle filters for mixed linear/nonlinear state-space models," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2279–2289, Jul. 2005.

[207] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Area. Commun.*, vol. 10, no. 5, pp. 819–829, 1992.

[208] L. Mitiche, B. Derras, and A. B. H. Adamou-Mitiche, "Speech modelling by model-order reduction: Snr behaviour," *IEEE Electr. Letters*, vol. 39, no. 17, pp. 1288–1290, Aug. 2003.

[209] L. Mitiche, A. B. H. Adamou-Mitiche, and D. Berkani, "Low-order model for speech signals," *Signal Proc.*, vol. 84, no. 10, pp. 1805–1811, Oct. 2004.

[210] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Language Process.*, vol. 7, 1998, pp. 2819–2822.

[211] G. Fant, *Acoustic theory of speech production*. Mouton de Gruyter, Jan. 1970.

[212] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[213] M. J. Daly, J. P. Reilly, and J. H. Manton, "A Bayesian approach to blind source recovery," in *Asil. Conf. Signals, Syst., Comput.*, vol. 1, Nov. 2004, pp. 989–993.

[214] J. J. Rajan, P. J. W. Rayner, and S. J. Godsill, "Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler," *IEE Proc.-Vis. Image Signal Process.*, vol. 144, no. 4, pp. 249–256,

August 1997.

[215] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, pp. 899–011, Aug. 1983.

[216] L. A. Liporace, "Linear estimation of nonstationary signals," *J. Acoust. Soc. Amer.*, vol. 58, no. 6, pp. 1288–1295, December 1976.

[217] R. Charbonnier, M. Barlaud, G. Alengrin, and J. Menez, "Results on AR-modelling of nonstationary signals," *Signal Proc.*, vol. 12, no. 2, pp. 143–151, Mar. 1987.

[218] J. J. Rajan and P. J. W. Rayner, "Generalized feature extraction for time-varying autoregressive models," *IEEE Trans. Signal Process.*, vol. 44, no. 10, pp. 2498–2507, Oct. 1996.

[219] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 387–392, Apr. 1985.

[220] A. P. Liavas, P. A. Regalia, and J.-P. Delmas, "Blind channel approximation: Effective channel order determination," *IEEE Trans. Signal Process.*, vol. 47, no. 12, pp. 3336–3344, Dec. 1999.

[221] H. Monson, *Statistical digital signal processing and modeling*. New Jersey, NJ: John Wiley & Sons, 1996.

[222] J. R. Hopgood, "Models for blind speech dereverberation: A subband all-pole filtered block stationary autoregressive process," in *Proc. EUSIPCO*, Antalya, Turkey, Sep. 2005.

[223] J. R. Hopgood and S. I. Hill, "An exact solution for incorporating boundary continuity constraints in subband all-pole modelling," in *Proc. IEEE Conf. SSP*, Jul. 2005, pp. 835–840.

[224] S. Theodoridis and K. Koutroumbas, *Pattern recognition*, 2nd ed. San Diego, CA: Academic Press, 2003.

[225] R. L. Bouqin-Jeannès and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Commun.*, vol. 16, no. 3, pp. 245–254, Apr. 1994.

[226] D. Koller and R. Fratkina, "Using learning for approximation in stochastic processes," in *Proc. Int. Conf. Machine Learning*, 1998.

[227] D. Fox, W. Burgart, F. Dellaert, and S. Thrun, "Monte Carlo localization: Efficient position estimation for mobile robots," in *Proc. Nat. Conf. Art. Intell.*, 1999.

[228] V. Smidl and A. P. Quinn, *The variational Bayes method in signal processing*, ser. Signals and communication technology. Birkhäuser, 2006.

[229] R. D. Koilpillai and P. P. Vaidyanathan, "Cosine-modulated FIR banks satisfying perfect reconstruction," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 770–783, Apr. 1992.

[230] P. P. Vaidyanathan, "Multirate digital filters, filter bankds, polyphase networks,

and applications: A tutorial," *Proc. IEEE*, vol. 78, no. 1, pp. 56–93, Jan. 1990.

[231] ——, *Multirate systems and filter banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[232] M. Vetterli and J. Kovacević, *Wavelets and subband coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[233] S. Weiss, L. Lampe, and R. W. Stewart, "Efficient subband adaptive filtering with oversampled GDFT filter banks," in *Dig. IEE. Colloq. Adapt. Signal Proc. for Mobile Commun. Syst.*, vol. 4, no. 383, Oct. 1997, pp. 1–9.

[234] S. Weiss, A. Stenger, R. W. Stewart, and R. Rabenstein, "Steady-state performance limitations of subband adaptive filters," *IEEE Trans. Signal Process.*, vol. 49, pp. 1982–1991, 2001.

[235] S. Weiss and R. W. Stewart, *On adaptive filtering in oversampled subbands*, B. Girod and J. Huber, Eds. Shaker Verlag, 1998, vol. 8.

[236] S. Weiss, L. Lampe, and R. W. Stewart, "Efficient subband adaptive filtering with oversampled GDFT filter banks," in *Proc. IEEE Conf. IWAENC*, 1997.

[237] B. Farhang-Boroujeny and Z. Wang, "Adaptive filtering in subbands: Design issues and experimental results for acoustic echo cancellation," *Signal. Proc.*, vol. 61, pp. 213–223, 1997.

[238] P. A. Naylor, O. Tanrikulu, and A. G. Constantinides, "Subband adaptive filtering for acoustic echo control using allpass polyphase IIR filterbanks," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 143–155, Mar. 1998.

[239] M. Bolić, P. M. Djurić, and S. Hong, "Resampling algorithms for particle filters: A computational complexity perspective," *EURASIP J. Appl. Sig. Proc.*, vol. 15, pp. 2267–2277, 2004.

[240] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. Cambridge, MA: MIT Press, 2001.

[241] K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: Experimental validation," *EURASIP J. Audio, Speech, Music Proc.*, 2007.

[242] M. J. Daly and J. R. Reilly, "Blind deconvolution using Bayesian methods with application to the dereverberation of speech," in *Proc. IEEE Conf. ICASSP*, vol. 2, May 2004, pp. 1009–1012.

[243] A. Gilloire and M. Vetterli, "Adaptive filtering in subbansd: Analysis, experimnts, and application to acoustic echo cancellation," *IEEE Trans. Signal Process.*, vol. 40, pp. 1862–1875, Aug. 1992.

[244] M. Harteneck, J. M. P. Borallo, and R. W. Stewart, "An oversampled subband adptive filter without cross adaptive filters," *Signal Proc.*, vol. 64, no. 1, pp. 93–101, 1998.

[245] M. Harteneck, S. Weiss, and R. W. Stewart, "Design of near perfect reconstruction oversampled filterbanks for subband adaptive filters," *IEEE Trans. Circuits*

Syst. II, vol. 46, pp. 1081–1085, Aug. 1999.

[246] J. P. Reilly, M. Wilbur, M. Seibert, and N. Ahmadvand, "The complex subband decomposition and its application to the decimation of large adaptive filtering problems," *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2730–2743, Nov. 2002.

[247] Y. Boers and J. N. Driessen, "Interacting multiple model particle filter," *IEE Proc. Radar, Sonar and Navig.*, vol. 150, no. 5, pp. 344–349, Oct. 2003.

[248] E. A. Lehmann, A. Johansson, and S. Nordholm, "Reverberation-time prediction method for room impulse responses simulated with the image-source model," in *Proc. IEEE Conf. WASPAA*, New Paltz, NY, Oct. 2007, pp. 159–162.

[249] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: A flexible Bayesian approach," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 4, pp. 1283–1295, May 2007.

[250] J. M. Turnbull, A. T. Sapeluk, and R. I. Damper, "A new method of pole-tracking with application to vowel andsemivowel recognition," in *Proc. IEEE Conf. ICASSP*, vol. 1, Glasgow, UK, May 1989, pp. 568–571.

[251] H. H. Robertson, "Approximate design of digital filters," *Technometrics*, vol. 7, no. 3, pp. 387–403, Aug. 1965.

[252] B. Gold and L. R. Rabiner, "Analysis of digital and analog formant synthesizers," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 1, pp. 81–94, Aug. 1968.

[253] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer handbook of speech processing*. New York, NY: Springer Verlag, 2008.

Author index

Driessen, J. N. 208
Dubois, C. 211
Dullerud, G. E. 49

Elko, G. W. 14, 15, 21
Eneman, K. 202
Ephraim, Y. 18
Evers, C. xx, 18, 22, 26, 27, 46, 174, 176, 178

Fant, G. 147
Farhang-Boroujeny, B. 194, 207
Faucon, G. 188, 191
Fitch, W. T. 33
Flanagan, J. L. 14, 15, 21, 27, 32–34, 36
Florencio, D. 19, 25, 194
Fong, W. 52, 56
Fox, D. 194
Fratkina, R. 194
Frey, A. R. 35
Fujisaki, H. 45
Furui, S. 31

Gannot, S. 15, 19, 84, 86
Gaubitch, N. 15, 25, 194, 202
Gaubitch, N. D. 14, 19, 25, 27
Gdoura, I. J. 38
Gelb, A. 6
Gersho, A. 134
Geweke, J. 90
Giedd, J. 33
Gilks, W. R. 101
Gillespie, B. W. 19, 25, 194
Gilloire, A. 207
Godsill, S. 49, 101, 128
Godsill, S. J. 7, 26, 41, 49, 52, 56, 90, 96, 101, 105, 125, 128, 174
Gold, B. 31, 225
Goodall, R. 105
Gordon, N. xxv, 6, 7, 83, 84, 99, 101, 207
Gordon, N. J. 7, 94
Goubran, R. 38
Gray, A. H. 35, 218, 219
Grenier, Y. 15, 27, 174, 181
Griebel, S. 24
Griebel, S. M. 24, 27
Gruber, P. 49, 128
Gustaffson, F. 98, 100
Gustafsson, F. 105, 125
Guzman, S. J. 65

Habets, E. 22, 27
Habets, E. A. P. 19
Hall, M. G. 47, 174, 181
Hammersley, J. M. 7
Hammond, J. K. 70
Hanauer, S. L. 151
Handschin, J. E. 97
Haneda, Y. 70, 76
Hansen, J. 139, 188
Hansen, J. H. L. 41
Harteneck, M. 207
Haykin, S. 53, 159
Hendeby, G. 105
Hess, W. J. 23
Hidaka, T. 73
Higuchi, T. 98
Hikichi, T. 21, 25, 27
Hill, S. I. 178, 194, 202
Hol, J. D. 98, 100
Holt, N. J. 65
Hon, H. 31
Honda, K. 33
Hong, S. xxv, 98, 100, 101, 196
Hopgood, J. R. xx, 17, 18, 22, 26, 27, 46, 67, 73, 74, 133, 174, 176, 178, 194, 202
Huang, X. D. 31
Huang, Y. 16, 27

Ingrbrestsen, R. B. 17
Isard, M. 94
Ishizaka, K. 33
Itakura, F. 38, 71

Jacobs, O. L. R. 6
James, H. M. 49
Jan, E. E. 15, 27
Jelinek, F. 31
Jenkins, G. M. 55, 61, 132
Johansen, A. M. 82, 94
Johansson, A. 208

Johnston, J. D. 14, 15, 21
Juang, B.-H. 21
Julier, S. J. 86
Juntunen, M. 50
Jurafsky, D. 31

Kabal, P. 16, 17
Kadirkamanathan, V. 105
Kailath, T. 176
Kaipio, J. P. 50
Kalman, R. E. 6, 83
Kanazawa, K. 94
Kaneda, Y. 70, 76
Kaneko, T. 33
Karlsson, R. 105
Kay, S. M. 4, 46
Kelly, J. L. 34
Kennedy, R. 17
Kinoshita, K. 21, 22, 27, 173
Kinsler, L. E. 35
Kirubarajan, T. 7
Kitawaki, N. 70
Klatt, D. H. 36
Knight, W. C. 4
Koike, T. 16
Koilpillai, R. D. 194
Koller, D. 94, 194
Kompis, M. 69
Kong, A. 96, 97, 102
Koutroumbas, K. 186
Kovacević, J. 194
Kumamoto, H. 96
Kuttruff, H. 3, 65, 67, 68, 73, 233, 234

Lall, S. G. 49
Lampe, L. 194, 207
Lebart, K. 19, 27
Lehmann, E. A. 208
Leiserson, C. E. 196
Letowski, T. R. 3
Lewis, F. L. 6
Li, P. 105
Li, X. R. 7
Liavas, A. P. 176
Libbey, B. 3, 65
Lin, X. S. 25, 194
Linggard, R. 33, 36, 37, 62
Liporace, L. A. 174, 181
Litovsky, R. Y. 65
Liu, J. S. 96–98, 100, 102, 105
Lochbaum, C. C. 34
Loizou, P. 38
Loizou, P. C. 42–44, 134, 188
Lyapunov, A. M. 49
Lyon, R. H. 16

MacDonald, A. D. 3
MacEachern, S. N. 105
Mahmoud, S. 38
Makhoul, J. 41
Makino, S. 70, 76
Malah, D. 18
Malladi, K. M. 49, 128
Malvar, H. 19, 25, 194
Manolakis, D. G. 52, 53, 59, 61, 159, 220, 222
Manton, J. H. 173, 194
Mariano, R. S. 97
Markel, J. D. 35, 218, 219
Martin, J. 31
Maskell, S. xxv, 7, 99, 101
Mason, D. 65
Masuda, S. 73
Matsudaira, M. 33
Maybeck, P. S. 6
Mayne, D. Q. 97
Mehta, V. 69
Menez, J. 174, 181
Meziane, A. 38
Mitiche, L. 134
Miyoshi, M. 19–22, 25, 27, 173, 215, 216
Monson, H. 176
Moody, M. 69, 234
Moonen, M. 58, 59, 202
Moore, J. B. 7, 49, 50
Morgan, N. 31
Morton, K. W. 7
Moulines, E. 98
Mourjopoulos, J. 70
Mourjopoulos, J. N. 69–71, 234

Murthy, P. Satyanarayana 24, 138, 139
Myatt, T. 19

Nábelek, A. K. 3, 65
Nakatani, T. 19–22, 27, 173, 215, 216
Naylor, P. A. 14, 19, 25, 27, 194
Nelson, A. T. 86
Nemer, E. 38
Nichols, N. B. 49
Niranjan, M. 41
Nishihara, N. 73
Nordholm, S. 208
Nordlund, P.-J. 125

Okano, T. 73
Oppenheim, A. V. 47, 174, 181
Orlikoff, R. F. 34
O'Shaughnessy, D. 31, 34, 43

Paraskevas, M. A. 70, 71
Pellom, B. 139, 188
Peterson, G. E. 21, 37
Petropulu, A. 17
Peutz, V. M. A. 3
Phillips, R. S. 49
Pintelon, R. 74
Prasanna, S. R. Mahadeva 24
Pridham, R. G. 4
Proakis, J. G. 41, 52, 53, 59, 61, 159, 220, 222

Quatieri, T. 43
Quinn, A. P. 194

Rabenstein, R. 194, 207
Rabiner, L. R. 18, 34, 36, 39–41, 53, 159, 222, 225
Radlović, B. 17
Rajakumar, R. V. 49, 128
Rajan, J. J. 174
Rao, K. Sreenivasa 24
Rao, T. S. 48
Rayner, P. J. W. 22, 74, 174
Regalia, P. A. 176
Reilly, J. P. 173, 194, 207
Reilly, J. R. 202
Reinsel, G. C. 55, 61, 132
Ristic, B. 6, 7, 83, 84, 101, 207
Rivest, R. L. 196
Robert, C. P. 7, 99, 103, 125
Roberts, G. O. 88
Robertson, H. H. 225
Rogers, P. H. 3, 65
Rosenberg, A. 42
Rubinstein, B. Y. 90
Russell, S. 94

Salmond, A. F. M 7, 94
Salmond, D. J. 7, 94
Sanders, J. V. 35
Sapeluk, A. T. 224
Schafer, R. W. 34, 36, 39–41, 53, 159, 222
Schmidt, S. F. 86
Schön, T. 125
Schon, T. B. 98, 100
Schoukens, J. 74
Seibert, M. 207
Sekey, A. 134
Shimelevich, L. I. 96, 129
Smidl, V. 194
Smith, A. F. M. 88
Spanias, A. 38
Spencer, P. S. 22, 71, 74, 75
Sridharan, S. 69, 234
Stein, C. 196
Stenger, A. 194, 207
Stevens, K. 31
Stewart, R. W. 194, 207
Stockham, T. G. Jr. 17
Storvik, G. 9
Subramaniam, S. 17
Surendran, A. C. 15, 27
Svetnik, V. B. 96, 129

Tachibana, H. 19, 20, 173, 215, 216
Takata, Y. 65
Takeda, K. 38
Tanizaki, H. 7, 97
Tanrikulu, O. 194
Taylor, W. K. 65
Teager, H. 43
Teager, S. 43

Tervo, J. 50
Theodoridis, S. 186
Therrien, C. W. 50, 53, 159
Thompson, C. 69
Thrun, S. 194
Todtli, J. 49, 128
Tohyama, M. 16
Trager, G. L. 32
Tran, H. D. 38
Trees, H. L. V. 18
Tucker, F. M. 3
Turnbull, J. M. 224

Uhlmann, J. K. 86

Vaidyanathan, P. P. 194, 207
van der Merwe, R. 86
Van Veen, B. D. 14, 15
van Waterschoot, T. 58, 59
Vermaak, J. 41, 49, 101, 125, 128
Vetterli, M. 194, 207

Wan, E. A. 86
Wang, D. 19
Wang, H. 71
Wang, S. 134
Wang, X. 7, 88
Wang, Z. 194, 207
Ward, D. 25
Ward, D. B. 25, 27
Watkins, A. J. 65
Wax, M. 176
Weinstein, E. 15
Weiss, S. 194, 207
Wen, J. Y. C. 19
Wendt, C. 17
West, M. 49, 52, 56, 101, 128
Wilbur, M. 207
Willsky, A. S. 47, 174, 181
Winther, O. 36, 56, 63, 151, 152, 161, 164–166
Wong, W. H. 96, 97, 102
Wu, M. 19

Yegnanarayana, B. 24, 138, 139
Yeredor, A. 84, 86
Yoshioka, T. 19–21, 173, 215, 216
Yost, W. A. 65

Zahn, R. 14, 15, 21
Zaritskii, V. S. 96, 129

Alice laughed. “There’s no use trying," she said. “One *can’t* believe impossible things."

“I daresay you haven’t had much practice," said the Queen. “When I was your age, I always did it for half-an-hour a day. Why, sometimes I’ve believed as many as six impossible things before breakfast. There goes the shawl again!’"

LEWIS CARROLL, *“Through the Looking Glass”* (1871)